BANCA D'ITALIA
EUROSISTEMA

Temi di Discussione

(Working Papers)

Evaluating students' evaluations of professors

by Michela Braga, Marco Paccagnella and Michele Pellizzari

October 2011

Number 825

BANCA D'ITALIA
EUROSISTEMA

Temi di discussione

(Working papers)

Evaluating students' evaluations of professors

by Michela Braga, Marco Paccagnella and Michele Pellizzari

Number 825 - October 2011

The purpose of the Temi di discussione *series is to promote the circulation of working papers prepared within the Bank of Italy or presented in Bank seminars by outside economists with the aim of stimulating comments and suggestions.*

The views expressed in the articles are those of the authors and do not involve the responsibility of the Bank.

EVALUATING STUDENTS' EVALUATIONS OF PROFESSORS

by Michela Braga*, Marco Paccagnella† and Michele Pellizzari‡

Abstract

This paper contrasts measures of teacher effectiveness with the students' evaluations of the same teachers using administrative data from Bocconi University (Italy). The effectiveness measures are estimated by comparing the subsequent performance in follow-on coursework of students who are randomly assigned to teachers in each of their compulsory courses. We find that, even in a setting where the syllabuses are fixed, teachers still matter substantially. Additionally, we find that our measure of teacher effectiveness is negatively correlated with the students' evaluations of professors: in other words, teachers who are associated with better subsequent performance receive worse evaluations from their students. We rationalize these results with a simple model where teachers can either engage in real teaching or in teaching-to-the-test, the former requiring greater student effort than the latter. Teaching-to-the-test guarantees high grades in the current course but does not improve future outcomes. Hence, if students are short-sighted and give better evaluations to teachers from whom they derive higher utility in a static framework, the model is capable of predicting our empirical finding that good teachers receive bad evaluations.

JEL Classification: I20.
Keywords: teacher quality, postsecondary education.

Contents

* University of Milan.

† Bank of Italy, Trento Branch, Economic Research Unit.

‡ Bocconi University, IGIER, IZA and C.F. Dondena Centre.

1 Introduction[1]

The use of anonymous students' evaluations of professors to measure teachers' performance has become extremely popular in many universities around the world (Becker and Watts, 1999). These normally include questions about the clarity of lectures, the logistics of the course, and many others. They are either administered to the students during a teaching session toward the end of the term or, more recently, filled on-line.

From the point of view of the university administration, such evaluations are used to solve the agency problems related to the selection and motivation of teachers, in a context in which neither the types of teachers, nor their levels of effort, can be observed precisely. In fact, students' evaluations are often used to inform hiring and promotion decisions (Becker and Watts, 1999) and, in institutions that put a strong emphasis on research, to avoid strategic behavior in the allocation of time or effort between teaching and research activities (Brown and Saks, 1987).[2]

The validity of anonymous students' evaluations as indicators of teacher ability rests on the assumption that students are in a better position to observe the performance of their teachers. While this might be true for the simple fact that students attend lectures, there are also many reasons to question the appropriateness of such a measure. For example, the students' objectives might be different from those of the principal, i.e. the university administration. Students may simply care about their grades, whereas the university (or parents or society as a whole) cares about their learning and the two (grades and learning) might not be perfectly correlated, especially when the same professor is engaged both in teaching and in grading the exams.

[1]We would like to thank Bocconi University for granting access to its administrative archives for this project. In particular, the following persons provided invaluable and generous help: Giacomo Carrai, Mariele Chirulli, Mariapia Chisari, Alessandro Ciarlo, Alessandra Gadioli, Roberto Grassi, Enrica Greggio, Gabriella Maggioni, Erika Palazzo, Giovanni Pavese, Cherubino Profeta, Alessandra Startari and Mariangela Vago. We are also indebted to Tito Boeri, Giovanni Bruno, Giacomo De Giorgi, Marco Leonardi, Tommaso Monacelli, Tommy Murphy and Tommaso Nannicini for their precious comments. We would also like to thank seminar participants at the Bank of Italy, Bocconi University, London School of Economics, UC Berkeley, Università Statale di Milano and LUISS University. Davide Malacrino and Alessandro Ferrari provided excellent research assistance. The views expressed in this paper are solely those of the authors and do not involve the responsibility of the Bank of Italy. The usual disclaimer applies. *Corresponding author*: Michele Pellizzari, Department of Economics, Bocconi University, via Roentgen 1, 20136 Milan - Italy; phone: +39 02 5836 3413; fax: +39 02 5836 3309; email: michele.pellizzari@unibocconi.it.

[2]Although there is some evidence that a more research oriented faculty also improve academic and labor market outcomes of graduate students (Hogan, 1981).

Consistently with this interpretation, Krautmann and Sander (1999) show that, conditional on learning, teachers who give higher grades also receive better evaluations, a finding that is confirmed by several other studies and that is thought to be a key cause of grade inflation (Carrell and West, 2010; Weinberg, Fleisher, and Hashimoto, 2009).

Measuring teaching quality is complicated also because the most common observable teachers' characteristics, such as their qualifications or experience, appear to be relatively unimportant (Hanushek and Rivkin, 2006; Krueger, 1999; Rivkin, Hanushek, and Kain, 2005). Despite such difficulties, there is also ample evidence that teachers' quality matters substantially in determining students' achievement (Carrell and West, 2010; Rivkin, Hanushek, and Kain, 2005) and that teachers respond to incentives (Duflo, Hanna, and Kremer, 2010; Figlio and Kenny, 2007; Lavy, 2009). Hence, understanding how professors should (or should not) be monitored and incentivized is of primary importance.

In this paper we evaluate the content of the students evaluations by contrasting them with objective measures of teacher effectiveness. We construct such measures by comparing the performance in subsequent coursework of students who are randomly allocated to different teachers in their compulsory courses. For this exercise we use data about one cohort of students at Bocconi University - the 1998/1999 freshmen - who were required to take a fixed sequence of compulsory courses and who where randomly allocated to a set of teachers for each of such courses. Additionally, the data are exceptionally rich in terms of observable characteristics, in particular they include measures of cognitive ability, family income and entry wages, which are obtained from regular surveys of graduates.[3]

We find that, even in a setting where the syllabuses are fixed and all teachers in the same course present exactly the same material, professors still matter substantially. The average difference in subsequent performance between students who were assigned to the best and worst teacher (on the effectiveness scale) is approximately 23% of a standard deviation in the distribution of exam grades, corresponding to about 3% of the average grade. This effect translates into approximately 1.4% of the average entry wage or 14 euros per month (160-200 euros per year). Moreover, our measure of teaching quality appears to be negatively correlated with the

[3]The same data are used in De Giorgi, Pellizzari, and Redaelli (2010).

students' evaluations of the professors: in other words, teachers who are associated with better subsequent performance receive worst evaluations from their students. On the other hand, teachers who are associated with high grades in their own exams receive better evaluations.

We rationalize these results with a simple model, where good teachers are those who provide their students with knowledge that is useful in future learning and, at the same time, require high effort from their students. Students are heterogeneous in their disutility of effort, which is higher for the least able ones, and evaluate professors on the basis of their realized utility, which depends on grades/learning and effort. In this setting, students in the bottom part of the ability distribution may, in fact, give worse evaluations to the good teachers, who impose a high effort cost on them, than the bad teachers.

Consistently with these predictions, we also find that the evaluations of classes in which high skill students (identified by their score in the cognitive admission test) are over-represented are more in line with the estimated real teacher quality. Furthermore, the distributions of grades in the classes of the most effective teachers are more dispersed, a piece of evidence that lends support to our specification of the learning function. Additionally, in order to support our assumption that evaluations are based on students' realized utility, we match our data with the weather conditions observed on the exact days when students filled the evaluation questionnaires. Under the assumption that the weather affects utility and not teaching quality, finding that the students' evaluations react to meteorological conditions lends support to the specification of our model.[4] Our results show that students evaluate professors more negatively on rainy and cold days.

There is a large literature that investigates the role of teacher quality and teacher incentives in improving educational outcomes, although most of the existing studies focus on primary and secondary schooling (Figlio and Kenny, 2007; Jacob and Lefgren, 2008; Kane and Staiger, 2008; Rivkin, Hanushek, and Kain, 2005; Rockoff, 2004; Rockoff and Speroni, 2010; Tyler, Taylor, Kane, and Wooten, 2010). The availability of standardized test scores facilitates the

[4]One may actually think that also the mood of the professors, hence, their effectiveness in teaching is affected by the weather. However, students' are asked to evaluate teachers' performance over the entire duration of the course and not exclusively on the day of the test. Moreover, it is a clear rule of the university to have students fill the questionnaires before the lecture, so that the teachers' performance on that specific day should not affect the evaluations.

evaluation of teachers in primary and secondary schools and such tests are currently available in many countries and also across countries (Mullis, Martin, Robitaille, and Foy, 2009; OECD, 2010). The large degree of heterogeneity in subjects and syllabuses in universities makes it very difficult to design common tests that would allow to compare the performance of students who were exposed to different teachers, especially across subjects. At the same time, the large increase in college enrollment experienced in almost all countries around the world in the past decades (OECD, 2008) calls for a specific focus on higher education, as in this study.[5]

To the best of our knowledge, only three other papers investigate the role of students' evaluations in university, namely Carrell and West (2010), Hoffman and Oreopoulos (2009) and Weinberg, Fleisher, and Hashimoto (2009). Compared to these papers we improve in various directions. First of all, the random allocation of students to teachers in our setting differentiates our approach from that of Hoffman and Oreopoulos (2009) and Weinberg, Fleisher, and Hashimoto (2009), who cannot purge their estimates from the potential bias due to the best students selecting the courses of the best professors. Rothstein (2009) and Rothstein (2010) show that correcting such a selection bias is pivotal to producing reliable measures of teaching quality.

The study of Carrell and West (2010), a paper that was developed parallelly and independently of ours, is perhaps the most similar to ours, both in terms of methodology and results. They also document a surprising negative correlation between the students' evaluations of professors and harder measures of teaching quality, however, we improve on their analysis in at least three important dimensions. First and most important, we provide a theoretical framework for the interpretation of such a striking finding, which is absent in Carrell and West (2010). Given that our results forcefully challenge the current most popular method used by most universities around the world to measure the teaching performance of their employees, it is paramount to provide a model that can rationalize the behaviors of both students and professors which generate the observed data. Furthermore, we show that our theory is consistent with additional pieces of evidence and we use it to formulate policy proposals.

[5]On average in the OECD countries 56% of school-leavers enrolled in tertiary education in 2006 versus 35% in 1995. The same secular trends appear in non-OECD countries. Further, the number of students enrolled in tertiary education has increased on average in the OECD countries by almost 20% between 1998 and 2006, with the US having experienced a higher than average increase from 13 to 17 millions.

Second, by observing wages for our students we are able to attach a price tag to our measures of teacher quality, something that, to our knowledge, has never been possible in previous studies. [6]

Finally, Carrell and West (2010) use data from a U.S. Air Force Academy, while our empirical application is based on a more standard institution of higher education.[7] In particular, the vast majority of the students in our sample enter a standard labor market when they graduate, whereas the cadets in Carrell and West (2010) are required to serve as officers in the U.S. Air Force for 5 years after graduation and many probably pursue a longer military career. There are many reasons why the behaviors of both teachers, students and the university/academy might vary depending on the labor market they face. For example, students may put particular effort on some exams or activities that are particularly important in the military setting - like physical activities - at the expenses of other subjects and teachers and administrators may do the same.

More generally, this paper is also related and contributes to the wider literature on performance measurement and performance pay. For example, one concern with the students' evaluations of teachers is that they might divert professors from activities that have a higher learning content for the students (but that are more demanding in terms of students' effort) and concentrate more on classroom entertainment (popularity contests) or change their grading policies. This interpretation is consistent with the view that teaching is a multi-tasking job, which makes the agency problem more difficult to solve (Holmstrom and Milgrom, 1994). Subjective evaluations, which have become more and more popular in modern human resource practices, can be seen as a mean to address such a problem and, given the very limited extant empirical evidence (Baker, Gibbons, and Murphy, 1994; Prendergast and Topel, 1996), our results can certainly inform also this area of the literature.

[6] Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011) present some results in this same spirit but in a very different context (kindergarten) and without explicitly looking at measures of teaching quality (they rather consider teachers' experience).

[7] Bocconi is a selective college that offers majors in the wide area of economics, management, public policy and law, hence it is likely comparable to US colleges in the mid-upper part of the quality distribution. For example, faculty in the economics department hold PhDs from Harvard, MIT, NYU, Stanford, UCLA, LSE, Pompeu Fabra, Stockholm University. Recent top Bocconi PhD graduates landed jobs (either tenure track positions or post-docs) at the World Bank and the University College of London. Also, the Bocconi Business school is normally ranked in the same range as the Georgetown University McDonough School of Business or the Johnson School at Cornell University in the US and to the Manchester Business School or the Warwick Business School in the UK (see the *Financial Times Business Schools Rankings*).

The paper is organized as follows. Section 2 describes our data and the institutional details of Bocconi University. Section 3 presents our strategy to estimate teacher effectiveness and shows the results. In Section 4 we correlate teacher effectiveness with the students' evaluations of professors. Robustness checks are reported in Section 5. In Section 6 we present a simple theoretical framework that rationalizes our results, while Section 7 discusses some additional evidence that corroborates our model. Finally, Section 8 concludes.

2 Data and institutional details

The empirical analysis in this paper is based on data for one enrollment cohort of undergraduate students at Bocconi university, an Italian private institution of tertiary education offering degree programs in economics, management, public policy and law. We select the cohort of the 1998/1999 freshmen for technical reasons, being the only one available in our data where students were randomly allocated to teaching classes for each of their compulsory courses.[8]

In later cohorts, the random allocation was repeated at the beginning of each academic year, so that students would take all the compulsory courses of each academic year with the same group of classmates, which only permits to identify the joint effectiveness of the entire set of teachers in each academic year.[9] For earlier cohorts the class identifiers, which are the crucial piece of information for our study, were not recorded in the university archives.

The students entering Bocconi in the 1998/1999 academic year were offered 7 different degree programs, although only three of them attracted a sufficient number of students to require the splitting of lectures into more than one class: Management, Economics and Law&Management[10]. Students in these programs were required to take a fixed sequence of compulsory courses that span the entire duration of their first two years, a good part of their third year and, in a few cases, also their last year. Table 1 lists the exact sequence for each of

[8]The terms *class* and *lecture* often have different meanings in different countries and sometimes also in different schools within the same country. In most British universities, for example, *lecture* indicates a teaching session where an instructor - typically a full faculty member - presents the main material of the course; *classes* are instead practical sessions where a teacher assistant solves problem sets and applied exercises with the students. At Bocconi there was no such distinction, meaning that the same randomly allocated groups were kept for both regular lectures and applied classes. Hence, in the remainder of the paper we use the two terms interchangeably.

[9]De Giorgi, Pellizzari, and Woolston (2011) use data for these later cohorts for a study of class size.

[10]The other degree programs were Economics and Social Disciplines, Economics and Finance, Economics and Public Administration.

the three programs that we consider, breaking down courses by the term (or semester) in which they were taught and by subject areas (classified with different colors: red for management, black for economics, green for quantitative subjects, blue for law).[11] In Section 3 we construct measures of teacher effectiveness for the professors of these compulsory courses. We do not consider elective subjects, as the endogenous self-selection of students would complicate the analysis.

Table 1: Structure of degree programs

	MANAGEMENT	ECONOMICS	LAW&MANAG.
Term I	Management I	Management I	Management I
	Private law	Private law	Mathematics
	Mathematics	Mathematics	
Term II	Microeconomics	Microeconomics	Accounting
	Public law	Public law	
	Accounting	Accounting	
Term III	Management II	Management II	Management II
	Macroeconomics	Macroeconomics	Statistics
	Statistics	Statistics	
Term IV	Business law	Financial mathematics	Accounting II
	Manag. of Public Administrations	Public economics	Fiscal law
	Financial mathematics	Business law	Financial mathematics
	Human resources management		
Term V	Banking	Econometrics	Corporate finance
	Corporate finance	Economic policy	
	Management of industrial firms		
Term VI	Marketing	Banking	
	Management III		
	Economic policy		
	Managerial accounting		
Term VII	Corporate strategy		
Term VIII			Business law II

The colors indicate the subject area the courses belong to: red=management, black=economics, green=quantitative, blue=law. Only compulsory courses are displayed.

Most (but not all) of the courses listed in Table 1 were taught in multiple classes (see Section 3 for details). The number of such classes varied across both degree programs and specific courses. For example, Management was the program that attracted the most students (over 70% in our cohort), who were normally divided into 8 to 10 classes. Economics and

[11] Notice that Economics and Management share exactly the same sequence of compulsory courses in the first three terms. Indeed, students in these two programs did attend these courses together and made a final decision about their major at the end of the third term. De Giorgi, Pellizzari, and Redaelli (2010) study precisely this choice. In the rest of the paper we abstract from this issue and we treat the two degree programs as entirely separated. In the Appendix we present some robustness checks to justify this approach (see Figure A-2).

Law&Management students were much fewer and were rarely allocated to more than just two classes. Moreover, the number of classes also varied within degree programs depending on the number of available teachers in each course. For instance, in 1998/99 Bocconi did not have a law department and all law professors were contracted from other nearby universities. Hence, the number of classes in law courses were normally fewer than in other subjects. Similarly, since the management department was (and still is) much larger than the economics or the mathematics department, courses in the management areas were normally split in more classes than courses in other subjects.

Regardless of the specific class to which students were allocated, they were all taught the same material. In other words, all professors of the same course were required to follow exactly the same syllabus, although some variations across degree programs were allowed (i.e. mathematics was taught slightly more formally to Economics students than Law&Management ones).

Additionally, the exam questions were also the same for all students (within degree program), regardless of their classes. Specifically, one of the teachers in each course (normally a senior person) acted as a coordinator, making sure that all classes progressed similarly during the term, defining changes in the syllabus and addressing specific problems that might have arisen. The coordinator also prepared the exam paper, which was administered to all classes. Grading was usually delegated to the individual teachers, each of them marking the papers of the students in his/her own class, typically with the help of one or more teaching assistants. Before communicating the marks to the students, the coordinator would check that there were no large discrepancies in the distributions across teachers. Other than this check, the grades were not curved, neither across nor within classes.

Table 2 reports some descriptive statistics that summarize the distributions of (compulsory) courses and their classes across terms and degree programs. For example, in the first term Management students took 3 courses, divided into a total of 24 different classes: management I, which was split into 10 classes; private law, 6 classes; mathematics, 8 classes. The table also reports basic statistics (means and standard deviations) for the size of these classes.

Our data cover in details the entire academic history of the students in these programs, including their basic demographics (gender, place of residence and place of birth), high school

Table 2: Descriptive statistics of degree programs

	Term							
Variable	I	II	III	IV	V	VI	VII	VIII
				Management				
No. Courses	3	3	3	4	3	4	1	-
No. Classes	24	21	23	26	23	27	12	-
Avg. Class Size	129.00	147.42	134.61	138.62	117.52	133.48	75.08	-
SD Class Size	73.13	80.57	57.46	100.06	16.64	46.20	11.89	-
				Economics				
No. Courses	3	3	3	3	2	1	-	-
No. Classes	24	21	23	4	2	2	-	-
Avg. Class Size	129.00	147.42	134.61	98.25	131.00	65.5	-	-
SD Class Size	73.13	80.57	57.46	37.81	0	37.81	-	-
				Law & Management				
No. Courses	3	4	4	4	2	-	-	1
No. Classes	5	5	5	6	3	-	-	1
Avg. Class Size	104.40	139.20	139.20	116.00	116.00	-	-	174.00
SD Class Size	39.11	47.65	47.67	44.96	50.47	-	-	0.00

leaving grades as well as the type of high school (academic or technical/vocational), the grades in each single exam they sat at Bocconi together with the date when the exams were sat. Graduation marks are observed for all non-dropout students.[12] Additionally, all students took a cognitive admission test as part of their application to the university and such test scores are available in our data for all the students. Moreover, since tuition fees varied with family income, this variable is also recorded in our dataset. Importantly, we also have access to the random class identifiers that allow us to identify in which class each students attended each of their courses.

Table 3 reports some descriptive statistics for the students in our data by degree program. The vast majority of them were enrolled in the Management program (74%), while Economics and Law&Management attracted 11% and 14%. Female students were generally under-represented in the student body (43% overall), apart from the degree program in Law&Management. About two thirds of the students came from outside the province of Milan, which is where Bocconi is located, and such a share increased to 75% in the Economics

[12]The dropout rate, defined as the number of students who, according to our data, do not appear to have completed their programs at Bocconi over the total size of the entering cohort, is just above 10%. Notice that some of these students might have transfered to another university or still be working towards the completion of their program, whose formal duration was 4 years. In Section 5 we perform a robustness check to show that excluding the dropouts from our calculations is irrelevant for our results.

Table 3: Descriptive statistics of students

Variable	Management	Economics	Law &Management	Total
1=female	0.408	0.427	0.523	0.427
1=outside Milan[a]	0.620	0.748	0.621	0.634
1=top Income Bracket[b]	0.239	0.153	0.368	0.248
1=academic high school[c]	0.779	0.794	0.684	0.767
1=late enrollee[d]	0.014	0.015	0.011	0.014
High-school grade (0-100)	86.152	93.053	88.084	87.181
	(10.905)	(8.878)	(10.852)	(10.904)
Entry Test Score (0-100)	60.422	63.127	58.894	60.496
	(13.069)	(15.096)	(12.262)	(13.224)
University Grades (0-30)	25.684	27.032	25.618	25.799
	(3.382)	(2.938)	(3.473)	(3.379)
Wage (Euro)[e]	966.191	1,012.241	958.381	967.964
	(260.145)	(265.089)	(198.437)	(250.367)
Number of students	901	131	174	1,206

[a] Dummy equal to one if the student's place of residence at the time of first enrollment is outside the province of Milan (which is where Bocconi university is located).
[b] Family income is recorded in brackets and the dummy is equal to one for students who report incomes in the top bracket, whose lower threshold is in the order of approximately 110,000 euros at current prices.
[c] Dummy equal to one if the student attended a academic high school, such as a lyceum, rather than professional or vocational schools.
[d] Dummy equal to one if the student enrolled at Bocconi after age 19.
[e] Nominal value at current (2010) prices. Based on 391 observations for Management, 36 observations for Economics, 94 observations for Law&Management, i.e. 521 observations overall.

program. Family income was recorded in brackets and one quarter of the students were in the top bracket, whose lower threshold was in the order of approximately 110,000 euros at current prices. Students from such a wealthy background were under-represented in the Economics program and over-represented in Law&Management. High school grades and entry test scores (both normalized on the scale 0-100) provide a measure of ability and suggest that Economics attracted the best students, a fact that is confirmed by looking at university grades, graduation marks and entry wages in the labor market.

Data on wages come from graduate surveys that we were able to match with the administrative records. Bocconi runs regular surveys of all alumni approximately one to one and a half years since graduation. These surveys contain a detailed set of questions on labor market experience, including employment status, occupation, and (for the employed) entry wages. As it is common with survey data, not all contacts were successful but we were still able to

match almost 60% of the students in our cohort, a relatively good response rate for surveys.[13] Two years after graduation, the employment rate for students that graduated in 2002 and 2003 (surveyed in 2004 and 2005, respectively) was around 92%; 35% of the non-employed were continuing education. For this reason, entry wages is the only measure of labor market success we look at.

Finally, we complement our dataset with students' evaluations of teachers. Towards the end of each term (typically in the last week), students in all classes were asked to fill an evaluation questionnaire during one lecture. Questionnaires are distributed at the beginning of the lecture, and students are given a fair amount of time to fill in the forms (15-20 minutes). The questions gathered students' opinions about various aspects of the teaching experience, including the clarity of the lectures, the logistics of the course, the availability of the professor and so on. For each item in the questionnaire, students answered on a scale from 0 (very negative) to 10 (very positive) or 1 to 5.

In order to allow students to evaluate their experience without fear of retaliation from the teachers at the exam, such questionnaires are anonymous and it is impossible to match the individual student with a specific evaluation of the teacher. One might be worried that, nonetheless, students tend to give higher valuations to professors they fear the most, in order to please them in some way. We are not overly concerned about this issue.The evaluation of a single student has little weight (average class size is above 100), and students should be aware they have little chances of influencing the average valuation of the class. Furthermore, students know that the exams will be the same for all classes and that a single professor can't tailor the exam according to the desires of his students (only the coordinator can do that, to some extent, but we control for coordinator status). Students might know that some professors are more prone to "trade" good grades with good evaluations (and this might partially explain the positive correlation between evaluations and contemporaneous grades); this is consistent with our model in which students evaluate professors on the basis of their perceived utility and it would be a further argument

[13]The response rates are highly correlated with gender, because of compulsory military service, and with the graduation year, given that Bocconi has improved substantially over time in its ability to track its graduates. Until the 1985 birth cohort, all Italian males were required to serve in the army for 10-12 months but were allowed to postpone the service if enrolled in full time education. For college students, it was customary to enroll right after graduation.

against using students' evaluations as measures of teacher "quality".

Each questionnaire reports the name of the course and the class identifier, so that we can attach average evaluations to each class in each course. Figure A-1 in the Appendix shows, as an example, the first page of the evaluation questionnaire used in the academic year 1998-1999.[14]

In Table 4 we present some descriptive statistics of the answers to the evaluation questionnaires. We concentrate on a limited set of items, which we consider to be the most informative and interesting, namely overall teaching quality, lecturing clarity, the teacher's ability to generate interest in the subject, the logistic of the course and workload. These are the same items that we analyze in more details in Section 4. The exact wording and scaling of the questions are reported in Table A-4 in the Appendix.

The average evaluation of overall teaching quality is around 7, with a relatively large standard deviation of 0.9 and minor variations across degree programs. Although differences are not statistically significant, professors in the Economics program seem to receive slightly better students' evaluations than their colleagues in Management and, even more, in Law&Management. The same ranking holds for the other measures of teaching quality, namely the clarity of lecturing and the ability to generate interest in the subject. Economics compares slightly worse to the other programs in terms of course logistics

Some of the evaluation items are, understandably, highly correlated. For example, the correlation coefficient between overall teaching quality and lecturing clarity is 0.89. The course logistics and the ability of the teacher in generating interest for the subject are slightly less strongly correlated with the core measures of teacher quality (around 0.5-0.6). Workload is the least correlated with any other item (all correlation coefficients are below 0.2). The full correlation matrix is reported in Table A-5 in the Appendix.

Additionally, in Table 4 we also report the mean and standard deviations of the number of collected questionnaires and the number of officially enrolled students in each of class. One might actually be worried that students may drop out of a class in response to the quality of the

[14]The questionnaires were changed slightly over time as new items were added and questions were slightly rephrased. We focus on a subset of questions that are consistent over the period under consideration.

Table 4: Descriptive statistics of students' evaluations

Variable	Management mean (std.dev.)	Economics mean (std.dev.)	Law&Manag. mean (std.dev.)	Total mean (std.dev.)
Overall teaching quality[a]	7.103	7.161	6.999	7.115
	(0.956)	(0.754)	(1.048)	(0.900)
Lecturing clarity[b]	3.772	3.810	3.683	3.779
	(0.476)	(0.423)	(0.599)	(0.467)
Teacher generates interest[a]	6.800	6.981	6.915	6.864
	(0.905)	(0.689)	(1.208)	(0.865)
Course logistic[b]	3.683	3.641	3.617	3.666
	(0.306)	(0.266)	(0.441)	(0.303)
Course workload[b]	2.709	2.630	2.887	2.695
	(0.461)	(0.542)	(0.518)	(0.493)
Questionnaires/students[c]	0.777	0.774	0.864	0.782
	(0.377)	(0.411)	(0.310)	(0.383)

[a] Scores range from 0 to 10.
[b] Scores range from 1 to 5.
[c] Number of collected valid questionnaires over the number of officially enrolled students.
See Table A-4 for the exact wording of the evaluation questions.

teaching so that at the end of the course, when questionnaires are distributed only the students who liked the teacher are eventually present. Such a process would lead to a compression of the distribution of the evaluations, with good teachers being evaluated by their entire class (or by a majority of their allocated students) and bad teachers being evaluated only by a subset of students who particularly liked them.

The descriptive statistics reported in Table 4 seem to indicate that this is not a major issue, as on average the number of collected questionnaires is around 80% of the total number of enrolled students (the median is very similar). Moreover, when we correlate our measures of teaching effectiveness with the evaluations we condition on the official size of the class and we weight observations by the number of questionnaires.

Indirectly, the relatively high number of questionnaires over students is evidence that attendance was also pretty high. An alternative measure of attendance can be extracted from a direct question of the evaluation forms which asks students what percentage of the lectures they attended. Such a self-reported measure of attendance is also around 80%.

2.1 The random allocation

In this section we present evidence that the random allocation of students into classes was successful. De Giorgi, Pellizzari, and Redaelli (2010) use data for the same cohort (although for a smaller set of courses and programs) and provide similar evidence.

The randomization was (and still is) performed via a simple random algorithm that assigned a class identifier to each student, who were then instructed to attend the lectures for the specific course in the class labeled with the same identifier. The university administration adopted the policy of repeating the randomization for each course with the explicit purpose of encouraging wide interactions among the students.

Table 5 reports test statistics derived from regressions of the observable students' characteristics (by column) on class dummies.The null hypothesis under consideration is the joint significance of the coefficients on the class dummies, which amounts to testing for the equality of the means of the observable variables across classes. Notice that these are very restrictive tests, as it is sufficient to have one unbalanced class to make the test fail. Results show that the F statistics are never particularly high. In most cases the null cannot be rejected at conventional significance levels. The only exception is residence from outside Milan, which is abnormally low in two Management groups. Four outlier groups in the Economics program (out of the 72 classes that we considered) also seem to have a particularly low presence of female students, while high school grades appear slightly lower than average in 3 classes of the same program. Overall, Table 5 suggests that the randomization was rather successful.

Testing the equality of means is not a sufficient test of randomization for continuous variables. Hence, in Figure 1 we compare the distributions of our measures of ability (high school grades and entry test scores) for the entire student body and for a randomly selected class in each program. The figure evidently shows that the distributions are extremely similar and formal Kolmogorov-Smirnov tests confirm the visual impression.

Even though students were randomly assigned to classes, one may still be concerned about teachers being selectively allocated to classes. Although no explicit random algorithm was used to assign professors to classes, for obvious organizational reasons that was (and still is) done in

Table 5: Randomness checks - Students

	Female	Academic High School[a]	High School Grade	Entry Test Score	Top Income Bracket[a]	Outside Milan	Late Enrollees[a]
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
			Management				
Test statistics:	χ^2	χ^2	F	F	χ^2	χ^2	χ^2
mean	0.489	0.482	0.497	0.393	0.500	0.311	0.642
median	0.466	0.483	0.559	0.290	0.512	0.241	0.702
minimum	0.049	0.055	0.012	0.004	0.037	0.000	0.025
maximum	0.994	0.949	0.991	0.944	0.947	0.824	0.970
P-value[b] (total number of tests is 20)							
<0.01	0	0	0	1	0	3	0
<0.05	1	0	1	1	2	6	1
			Economics				
Test statistics:	χ^2	χ^2	F	F	χ^2	χ^2	χ^2
mean	0.376	0.662	0.323	0.499	0.634	0.632	0.846
median	0.292	0.715	0.241	0.601	0.616	0.643	0.911
minimum	0.006	0.077	0.000	0.011	0.280	0.228	0.355
maximum	0.950	0.993	0.918	0.989	0.989	0.944	0.991
P-value[b] (total number of tests is 11)							
<0.01	1	0	2	0	0	0	0
<0.05	1	0	2	1	0	0	0
			Law & Management				
Test statistics:	χ^2	χ^2	F	F	χ^2	χ^2	χ^2
mean	0.321	0.507	0.636	0.570	0.545	0.566	0.948
median	0.234	0.341	0.730	0.631	0.586	0.533	0.948
minimum	0.022	0.168	0.145	0.182	0.291	0.138	0.935
maximum	0.972	0.966	0.977	0.847	0.999	0.880	0.961
P-value[b] (total number of tests is 7)							
<0.01	0	0	0	0	0	0	0
<0.05	2	0	0	0	0	0	0

The reported statistics are derived from probit (columns 1,2,5,6,7) or OLS (columns 3 and 4) regressions of the observable students' characteristics (by column) on class dummies for each course in each degree program that we consider (Management: 20 courses, 144 classes; Economics: 11 courses, 72 classes; Law & Management: 7 courses, 14 classes). The reported p-values refer to tests of the null hypothesis that the coefficients on all the class dummies in each model are all jointly equal to zero. The test statistics are either χ^2 (columns 1,2,5,6,7) or F (columns 3 and 4), with varying parameters depending on the model.

[a] See notes to Table 3.

[b] Number of courses for which the p-value of the test of joint significance of the class dummies is below 0.05 or 0.01.

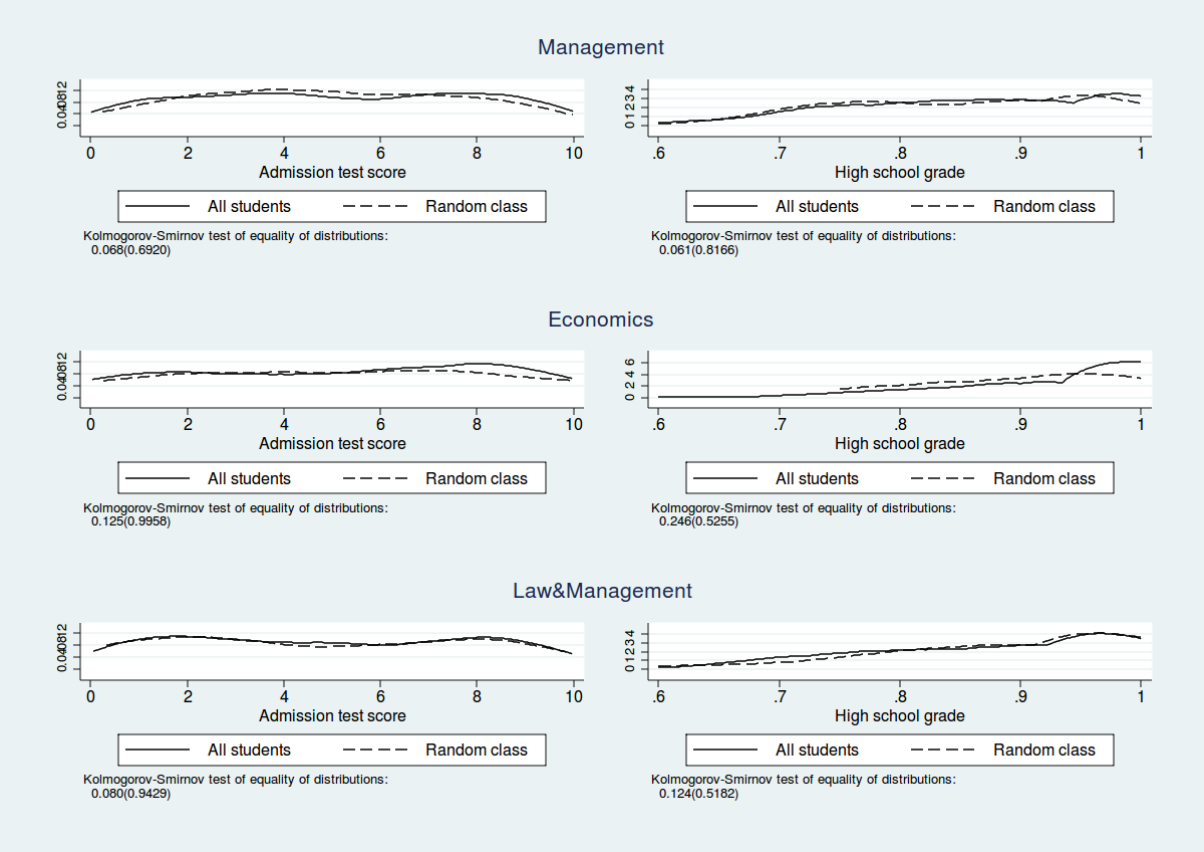


Figure 1: Evidence of random allocation - Ability variables

the Spring of the previous academic year, i.e. well before students were allowed to enroll, so that even if teachers were allowed to choose their class identifiers they would have no chance to know in advance the characteristics of the students who would be given that same identifier.

More specifically, there used to be (and still is) a very strong hysteresis in the matching of professors to class identifiers, so that, if no particular changes occurred, one kept the same class identifier of the previous academic year. It is only when some teachers needed to be replaced or the overall number of classes changed that modifications took place. Even in these instances, though, the distribution of class identifiers across professors changed only marginally. For example if one teacher dropped out, then a new teacher would take his/her class identifier and none of the others were given a different one. Similarly, if the total number of classes needed to be increases, the new classes would be added at the bottom of the list of identifiers with new teachers and no change would affect the existing classes and professors.[15]

[15] As far as we know, the total number of classes for a course has never been reduced.

About around the same time when teachers were given class identifiers (i.e. in the Spring of the previous academic year), also classrooms and time schedules were defined. On these two items, though, teachers did have some limited choice. Typically, the administration suggested a time schedule and room allocation and professors could request one or more modifications, which were accommodated only if compatible with the overall teaching schedule (e.g. a room of the required size was available at the new requested time).

In order to avoid any distortion in our estimates of teaching effectiveness due to the more or less convenient teaching times, we collected detailed information about the exact schedule of the lectures in all the classes that we consider, so that we can hold this specific factor constant (see Section 3). Additionally, we also know in which exact room each class was taught and we further conditions on the characteristics of the classrooms, namely the building and the floor where they are located. There is no variation in other features of the rooms, such as the furniture (all rooms were - and still are - fitted with exactly the same equipment: projector, computer, white-board) or the orientation (all rooms face the inner part of the campus where there is very limited car traffic).[16]

Table 6 provides evidence of the lack of correlation between teachers and classes' characteristics, namely we show the results of regressions of teachers' observable characteristics on classes' observable characteristics. For this purpose, we estimate a system of 9 seemingly unrelated simultaneous equations, where each observation is a class in a compulsory course. The dependent variables are 9 teachers' characteristics (age, gender, h-index, average citations per year and 4 dummies for academic positions) and the regressors are the class characteristics listed in the rows of the table.[17] The reported statistics test the null hypothesis that the coefficients on each class characteristic are all jointly equal to zero in all the equations of the system.[18] Results show that only the time of the lectures is significantly correlated with the teachers' observables at conventional statistical levels. In fact, this is one of the few ele-

[16] In principle we could also condition on room fixed effects but there are several rooms in which only one class of the courses that we consider was taught.

[17] The h-index is a quality-adjusted measure of individual citations based on search results on Google Scholar. It was proposed by Hirsch (2005) and it is defined as follows: *A scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have no more than h citations each.*

[18] To construct the tests we use the small sample estimate of the variance-covariance matrix of the system.

Table 6: Randomness checks - Teachers

	F-test	P-value
Class size[a]	0.94	0.491
Attendance[b]	0.95	0.484
Avg. high school grade	0.73	0.678
Avg. entry test score	1.37	0.197
Share of females	1.05	0.398
Share of students from outside Milan[c]	0.25	0.987
Share of top-income students[c]	1.31	0.228
Share academic high school[c]	1.35	0.206
Share late enrollees[c]	0.82	0.597
Share of high ability[d]	0.69	0.716
Morning lectures [e]	5.24	0.000
Evening lectures [f]	1.97	0.039
Room's floor[g]	0.45	0.998
Room's building[h]	1.39	0.188

The reported statistics are derived from a system of 9 seemingly unrelated simultaneous equations, where each observation is a class in a compulsory course (184 observations in total). The dependent variables are 9 teachers' characteristics (age, gender, h-index, average citations per year and 4 dummies for academic positions) and the regressors are the class characteristics listed in the table. The reported statistics test the null hypothesis that the coefficients on each class characteristic are all jointly equal to zero in all the equations of the system. The last row tests the hypothesis that the coefficients on all regressors are all jointly zero in all equations. All tests are distributed according to a F-distribution with (9,1467) degrees of freedom, apart from the joint test in the last row, which has (108,1467) degrees of freedom.

[a] Number or officially enrolled students.

[b] Attendance is monitored by random visits of university attendants to the class.

[c] See notes to Table 3.

[d] Share of students in the top 25% of the entry test score distribution.

[e] Share of lectures taught between 8.30 and 10.30 a.m.

[f] Share of lectures taught between 4.30 and 6.30 p.m.

[g] Test of the joint significance of 4 floor dummies.

[h] Dummy for building A.

ments of the teaching planning over which teachers had some limited choice. More specifically, professors are given a suggested time schedule for their classes in the spring of the previous academic year (usually based on the schedule of the current year), and they can either approve it or request changes. The administration, then, accommodates such changes only if they are compatible with the other many constraints in terms of rooms availability and course overlappings. In our empirical analysis we do control for all the factors in Table 6, so that our measures of teaching effectiveness are purged from the potential confounding effect of teaching times on students' learning.

3 Estimating teacher effectiveness

We use performance data for our students to estimate measures of teacher effectiveness. Namely, for each of the compulsory courses listed in Table 1 we compare the future outcomes of students that attended those courses in different classes, under the assumption that students who were taught by better professors enjoyed better outcomes later on. This approach is similar to the *value-added* methodology that is more commonly used in primary and secondary schools (Goldhaber and Hansen, 2010; Hanushek, 1979; Hanushek and Rivkin, 2006, 2010; Rivkin, Hanushek, and Kain, 2005; Rothstein, 2009) but it departs from its standard version, that uses contemporaneous outcomes and conditions on past performance, since we use future performance to infer current teaching quality.[19]

One most obvious concern with the estimation of teacher quality is the non-random assignment of students to professors. For example, if the best students self-select themselves into the classes of the best teachers, then estimates of teacher quality would be biased upward. Rothstein (2009) shows that such a bias can be substantial even in well-specified models and especially when selection is mostly driven by unobservabes.

We avoid these complications by exploiting the random allocation of students in our cohort to different classes for each of their compulsory courses. For this same reason, we focus exclusively on compulsory courses, as self-selection is an obvious concern for electives. Moreover, elective courses were usually taken by fewer students than compulsory ones and they were usually taught in one single class.

We compute our measures of teacher effectiveness in two steps. First, we estimate the conditional mean of the future grades (in compulsory courses) of students in each class according to the following procedure. Consider a set of students enrolled in degree program d and indexed by $i = 1, \ldots, N_d$, where N_d is the total number of students in the program. In our application there are three degree programs ($d = \{1, 2, 3\}$): Management, Economics and Law&Management. Each student i attends a fixed sequence of compulsory courses indexed by $c = 1, \ldots, C_d$, where C_d is the total number of such compulsory courses in degree program d. In each course c the student is randomly allocated to a class $s = 1, \ldots, S_c$, where S_c is the total

[19] For this reason we prefer to use the label *teacher effectiveness* for our estimates.

number of classes in course c. Denote by $\zeta \in Z_c$ a generic (compulsory) course, different from c, which student i attends in semester $t \geq t_c$, where t_c denotes the semester in which course c is taught. Z_c is the set of compulsory courses taught in any term $t \geq t_c$.

Let $y_{ids\zeta}$ denote the grade obtained by student i in course ζ. To control for differences in the distribution of grades across courses, $y_{ids\zeta}$ is standardized at the course level. Then, for each course c in each program d we run the following regression:

$$y_{ids\zeta} = \alpha_{dcs} + \beta X_i + \epsilon_{ids\zeta} \tag{1}$$

where X_i is a vector of student-level characteristics including a gender dummy, a dummy for whether the student is in the top income bracket, the entry test score and the high school leaving grade. The α's are our parameters of interest and they measure the conditional means of the future grades of students in class s: high values of α indicate that, on average, students attending course c in class s performed better (in subsequent courses) than students taking course c in a different class. The random allocation procedure guarantees that the class fixed effects α_{dcs} in equation 1 are purely exogenous and identification is straightforward.[20]

Notice that, since in general there are several subsequent courses ζ for each course c, each student is observed multiple times and the error terms $\epsilon_{ids\zeta}$ are serially correlated within i and across ζ. We address this issue by adopting a standard random effect model to estimate all the equations 1 (we estimate one such equation for each course c). Moreover, we further allow for cross-sectional correlation among the error terms of students in the same class by clustering the standard errors at the class level.

More formally, we assume that the error term is composed of three additive components (all with mean equal zero):

$$\epsilon_{ids\zeta} = \upsilon_i + \omega_s + \nu_{ids\zeta} \tag{2}$$

where υ_i and ω_s are, respectively, an individual and a class component, and $\nu_{ids\zeta}$ is a purely random term. Operatively, we first apply the standard random effect transformation to the

[20]Notice that in few cases more than one teacher taught in the same class, so that our class effects capture the overall effectiveness of teaching and cannot be attached to a specific person. Since the students' evaluations are also available at the class level and not for specific teachers, we cannot disaggregate further.

original model of equation 1.[21]

In the absence of other sources of serial correlation (i.e if the variance of ω_s were zero), such a transformation would lead to a serially uncorrelated and homoskedastic variance-covariance matrix of the error terms, so that the standard random effect estimator could be produced by running simple OLS on the transformed model. In our specific case, we further cluster the transformed errors at the class level to account for the additional serial correlation induced by the term ω_s.

Overall, we are able to estimate 230 such fixed effects, the large majority of which are for Management courses.[22]. Descriptive statistics of the estimated α's are reported in Table A-1 in the Appendix.

The second step of our approach is meant to purge the estimated α's from the effect of other class characteristics that might affect the performance of students in later courses but are not attributable to teachers. By definition, the class fixed effects capture all those features, both observable and unobservable, that are fixed for all students in the class. These certainly include teaching quality but also other factors that are documented to be important ingredients of the education production function, such as class size and class composition (De Giorgi, Pellizzari, and Woolston, 2011).

A key advantage of our data is that most of these other factors are observable. In particular, based on our academic records we can construct measures of both class size and class composition (in terms of students' characteristics). Additionally, we also have access to the identifiers of the teachers in each class and we can recover a large set of variables like gender, tenure status, and measures of research output. We also know which of the several teachers in each course acted as coordinator. These are the same teacher characteristics that we used in Table

[21]The standard random effect transformation subtracts from each variable in the model (both the dependent and each of the regressors) its within-mean scaled by the factor $\theta = 1 - \sqrt{\frac{\sigma_v^2}{|Z_c|(\sigma_\omega^2+\sigma_\nu^2)+\sigma_v^2}}$, where $|Z_c|$ is the cardinality of Z_c. For example, the random-effects transformed dependent variable is $y_{ids\zeta} - \theta\overline{y}_{ids}$, where $\overline{y}_{ids} = |Z_c|^{-1}\sum_{h=1}^{|Z_c|} y_{idh\zeta}$. Similarly for all the regressors. The estimates of σ_v^2 and $(\sigma_\omega^2 + \sigma_\nu^2)$ that we use for this transformation are the usual Swamy-Arora, also used by the command *xtreg* in Stata (Swamy and Arora, 1972).

[22]We cannot run equation 1 for courses that have no contemporaneous nor subsequent courses, such as Corporate Strategy for Management, Banking for Economics and Business Law for Law&Management (see Table 1). For such courses, the set Z_c is empty. Additionally, some courses in Economics and in Law&Management are taught in one single class, for example Econometrics (for Economics students) or Statistics (for Law&Management). For such courses, we have $S_c = 1$. The evidence that we reported in Tables 5 and 6 also refer to the same set of 230 classes.

6. Once we condition on all these observable controls, unobservable teaching quality is likely to be the only remaining factor that generates variation in the estimated α's. At a minimum, it should be uncontroversial that teaching quality is by far the single most important unobservable that generates variation in the $\widehat{\alpha}$'s, once conditioning on the observables.

The effect of social interactions among the students might also affect the estimated $\widehat{\alpha}$'s. However, notice that if such effects are related to the observable characteristics of the students, then we are able to control for those. Additionally, there might be complementarities among teachers' ability and students' interactions, as good teachers are also those who stimulate fruitful collaborations among their students. This component of the social interaction effects is certainly something that one would like to incorporate in a measure of teaching quality, as in our analysis.

Thus, in Table 7 we regress the estimated α's on all observable class and teacher characteristics. In column 1 we condition only on class size and class composition, in column 2 only on teachers' characteristics and in column 3 we combine the two sets of controls. In all cases we weight observations by the inverse of the standard error of the estimated α's to take into account differences in the precision of such estimates. Consistently with previous studies on the same data (De Giorgi, Pellizzari, and Woolston, 2011), we find that larger classes tend to be associated with worse learning outcomes, that classes with more able students, measured with either high school grades or the entry test score, also perform better and that a high concentration of high income students appears to be detrimental for learning. Overall, observable class characteristics explain about 8% of the variation in the estimated α's within degree program, term and subject cells, where subjects are defined as in Table 1.[23]

The results in column 2 show a non linear relationship between teachers' age and teaching outcomes, which might be rationalized with increasing returns to experience. Also, professors who are more productive in research seem to be less effective as teachers, when output is measured with the h-index. The effect is reversed using yearly citations but it never reaches acceptable levels of statistical significance. Finally, and consistently with the age effect, also

[23]The Partial R-squared reported at the bottom of the table refer to the R-squared of a partitioned regression where the dummies for the degree program, the term and the subject are partialled out.

Table 7: Determinants of class effects

Dependent variable = $\widehat{\alpha}_s$	[1]	[2][a]	[3]
Class size[b]	-0.000**	-	-0.000**
	(0.000)		(0.000)
Avg. HS grade	2.159**	-	2.360**
	(1.039)		(1.070)
Avg. entry test score	-1.140	-	-1.530
	(1.392)		(1.405)
Share of females	0.006	-	-0.094
	(0.237)		(0.245)
Share from outside Milan	-0.080	-	-0.078
	(0.203)		(0.201)
Share of top income[b]	-0.283	-	-0.331
	(0.271)		(0.278)
Share from academic HS	0.059	-	-0.054
	(0.301)		(0.313)
Share of late enrollees	-0.365	-	0.017
	(0.827)		(0.843)
Share of high ability[b]	0.733*	-	0.763*
	(0.394)		(0.390)
Morning lectures[b]	0.015	-	-0.015
	(0.037)		(0.040)
Evening lectures[b]	-0.175	-	-0.170
	(0.452)		(0.490)
1=coordinator	-	0.013	0.039
		(0.038)	(0.041)
Male	-	-0.017	-0.014
		(0.024)	(0.025)
Age	-	-0.013***	-0.013**
		(0.005)	(0.005)
Age squared	-	0.000**	0.000*
		(0.000)	(0.000)
H-index	-	-0.008	-0.007
		(0.006)	(0.006)
Citations per year	-	0.000	0.000
		(0.001)	(0.001)
Full professor[c]		0.116*	0.121*
		(0.066)	(0.072)
Associate professor[c]		0.113*	0.118*
		(0.062)	(0.067)
Assistant professor[c]		0.109*	0.123*
		(0.061)	(0.065)
Classroom characteristics[d]	yes	no	yes
Degree program dummies	yes	yes	yes
Subject area dummies	yes	yes	yes
Term dummies	yes	yes	yes
Partial R squared[e]	0.089	0.081	0.158
Observations	230	230	230

Observations are weighted by the inverse of the standard error of the estimated α's. * p<0.1, ** p<0.05,***p<0.01

[a] Weighted averages of individual characteristics if there is more than one teacher per class.

[b] See notes to Table 6.

[c] All variables regarding the academic position refer to the main teacher of the class. The excluded dummy is a residual category (visiting prof., external experts, collaborators.)

[d] Four floor dummies, one building dummy and a dummy for multi-classrooms classes.

[e] R squared computed once program, term and subject fixed effects are partialled out.

the professor's academic position matters, with a ranking that gradually improves from assistant to associate to full professors (other academic positions, such as external or non tenured-track teachers, are the excluded group). However, as in Hanushek and Rivkin (2006) and Krueger (1999), we find that the individual traits of the teachers explain less than a tenth of the (residual) variation in students' achievement. Overall, the complete set of observable class and teachers' variables explains approximately 15% of the (residual) variation.

Our final measures of teacher effectiveness are the residuals of the regression of the estimated α's on all the observable variables, i.e the regression reported in column 3 of Table 7. In Table 8 we present descriptive statistics of such measures.

Table 8: Descriptive statistics of estimated teacher effectiveness

	Management	Economics	Law & Management	Total
PANEL A: Std. dev. of estimated teacher effect				
mean	0.069	0.159	0.019	0.086
minimum	0.041	0.030	0.010	0.010
maximum	0.106	0.241	0.030	0.241
PANEL B: Largest minus smallest class effect				
mean	0.190	0.432	0.027	0.230
minimum	0.123	0.042	0.014	0.014
maximum	0.287	0.793	0.043	0.043
No. of courses	20	11	7	38
No. of classes	144	72	14	230

Teacher effectiveness is estimated by regressing the estimated class effects (α) on observable class and teacher's characteristics (see Table 7).

The overall standard deviation of teacher effectiveness is 0.086.[24] This average is the composition of a larger variation among the courses of the program in Economics (0.159) and a more limited variation in Management (0.069) and Law & Management (0.019). Recall that grades are normalized so that the distributions of the class effects are comparable across courses. Hence, these results can be directly interpreted in terms of changes in outcomes. In other words, the overall effect of increasing teacher effectiveness by one standard deviation is

[24]The standard deviation that we consider is the OLS estimate of the residuals of the regression in column 3 of Table 7

an increase in the average grade of subsequent courses by 0.086 standard deviations, roughly 0.3 of a grade point or 1.1% over the average grade of approximately 26.[25] Given an estimated conditional elasticity of entry wages to GPA of 0.45, such an effect would cost students slightly more than 0.5% of their average entry monthly wage of 967 euros, or about 60 euros per year.[26] Since in our data we only observe entry wages, it might well be that the long term effects of teaching quality are even larger.

In Table 8 we also report the standard deviations of teacher effectiveness of the courses with the least and the most variation to show that there is substantial heterogeneity across courses. Overall, we find that in the course with the highest variation (management I in the Economics program) the standard deviation of our measure of effectiveness is approximately a quarter of a standard deviation in grades. This compares to a standard deviation of essentially zero (0.010) in the course with the lowest variation (mathematics in the Law&Management program).

In the lower panel of Table 8 we show the mean (across courses) of the difference between the largest and the smallest indicators of teacher effectiveness, which allows us to compute the effect of attending a course in the class of the best versus the worst teacher. On average, this effect amounts to 0.230 of a standard deviation, that is almost 0.8 grade points or 3% over the average grade. As already noted above, this average effect masks a large degree of heterogeneity across subjects ranging from almost 80% to a mere 4% of a standard deviation.

To further understand the importance of these effects, we can also compare particularly lucky students, who are assigned to good teachers (defined as those in the top 5% of the distribution of effectiveness) throughout their sequence of compulsory courses, to particularly unlucky students, who are always assigned to bad teachers (defined as those in the bottom 5% of the distribution of effectiveness). The average grades of these two groups of students are 1.8 grade points apart, corresponding to over 7% of the average grade. Based on our estimate of the wage elasticity, this difference translates into a sizable 300-400 euros per year (30.45

[25] In Italy, university exams are graded on a scale 0 to 30, with pass equal to 18. Such a peculiar grading scale comes from historical legacy: while in primary, middle and high school students were graded by one teacher per subject on a scale 0 to 10 (pass equal to 6), at university each exam was supposed to be evaluated by a commission of three professors, each grading on the same 0-10 scale, the final mark being the sum of these three. Hence, 18 is pass and 30 is full marks. Apart from the scaling, the actual grading at Bocconi is performed as in the average US or UK university.

[26] In Italy wages are normally paid either 13 or 14 times over the year, once every month plus one additional payment around mid December (*tredicesima*) and around mid June (*quattordicesima*).

euros/month) or 3.15% over the average.

For robustness and comparison, we estimate the class effects in two alternative ways. First, we restrict the set Z_c to courses belonging to the same subject area of course c, under the assumption that good teaching in one course is likely to have a stronger effect on learning in courses of the same subject areas (e.g. a good basic mathematics teacher is more effective in improving students performance in financial mathematics than in business law). The subject areas are defined by the colors in Table 1 and correspond to the department that was responsible for the organization and teaching of the course. We label these estimates *subject* effects. Given the more restrictive definition of Z_c we can only produce these estimates for a smaller set of courses and using fewer observation, which is the reason why we do not take them as our benchmark.

Next, rather than using performance in subsequent courses, we run equation 1 with the grade in the same course c as the dependent variable. We label these estimates *contemporaneous* effects[27]. We do not consider these contemporaneous effects as alternative and equivalent measures of teacher effectiveness, but we will use them to show that they correlate very differently with the students' evaluations. Descriptive statistics for the subject and contemporaneous effects are reported in Tables A-3 and A-2 in the Appendix.

Table 9: Comparison of benchmark, subject and contemporaneous teacher effects

Dependent variable: Benchmark teacher effectiveness		
Subject	0.048**	-
	(0.023)	
Contemporaneous	-	-0.096***
		(0.019)
Program fixed effects	yes	yes
Term fixed effects	yes	yes
Subject fixed effects	yes	yes
Observations	212	230

Bootstrapped standard errors in parentheses. Observations are weighted by the inverse of the standard error of the dependent variable. * p<0.1, ** p<0.05,***p<0.01

In Table 9 we investigate the correlation between these alternative estimates of teacher ef-

[27] When estimating *contemporaneous* effects we include past grades in the vector of student-level characteristics of equation 1

fectiveness. Specifically, we report results from weighted OLS regressions with our benchmark estimates as the dependent variable and, in turn, the subject and the contemporaneous effects on the right hand side, together with dummies for degree program, term and subject area.[28]

Reassuringly, the subject effects are positively and significantly correlated with our benchmark, while the contemporaneous effects are negatively and significantly correlated with our benchmark, a result that is consistent with previous findings (Carrell and West, 2010; Krautmann and Sander, 1999; Weinberg, Fleisher, and Hashimoto, 2009) and to which we will return in Section 4.

4 Correlating teacher effectiveness and student evaluations

In this section we investigate the relationship between our measures of teaching effectiveness from Section 3 and the evaluations teachers receives from their students. We concentrate on two core items from the evaluation questionnaires, namely overall teaching quality and the overall clarity of the lectures. Additionally, we also look at other items: the teacher's ability in generating interest for the subject, the logistics of the course (schedule of classes, combinations of practical sessions and traditional lectures) and the total workload compared to other courses.

Formally, we estimate the following equation:

$$q^k_{dtcs} = \lambda_0 + \lambda_1 \widehat{\alpha}_{dtcs} + \lambda_2 C_{dtcs} + \lambda_3 T_{dtcs} + \gamma_d + \delta_t + \upsilon_c + \epsilon_{dtcs} \tag{3}$$

where q^k_{dtcs} is the average answer to question k in class s of course c in the degree program d (which is taught in term t), $\widehat{\alpha}_{dtcs}$ is the estimated class fixed effect from equation 1, C_{dtcs} is the set of class characteristics, T_{dtcs} is the set of teacher characteristics. γ_d, δ_t and υ_c are fixed effects for degree program, term and subject areas, respectively. ϵ_{dtcs} is a residual error term.

Notice that the class and teacher characteristics are exactly the same as in Table 7, so that equation 3 is equivalent to a partitioned regression model of the evaluations q_{dtcs} on our measures of teacher effectiveness, i.e. the residuals of the regressions in Table 7, where all the

[28] To take into account the additional noise due to the presence of generated regressors on the right hand side of these models, the standard errors are bootstrapped. Further, each observation is weighted by the inverse of the standard error of the dependent variable, which is also a generated variable.

observables and the fixed effects are partialled out.

Since the dependent variable in equation 3 is an average, we use weighted OLS, where each observation is weighted by the square root of the number of collected questionnaires in the class, which corresponds to the size of the sample over which the average answers are taken. Additionally, we also bootstrap the standard errors to take into account the presence of generated regressors (the $\widehat{\alpha}$'s).

The first four columns of Table 10 reports the estimates of equation 3 for a first set of core evaluation items, namely overall teaching quality and lecturing clarity. For each of these items we show results obtained using our benchmark estimates of teacher effectiveness and those obtained using the contemporaneous class effects.

Results show that our benchmark class effects are negatively associated with all the items that we consider. In other words, teachers who are more effective in promoting future performance receive worst evaluations from their students. This relationship is statistically significant for all items (but logistics and workload, which are features of the course that are common to all classes and over which individual teachers have little or no control), and is of sizable magnitude. For example, one standard deviation increase in teacher effectiveness reduces the students evaluations of overall teaching quality by about 50% of a standard deviation. Such an effect could move a teacher who would otherwise receive a median evaluation down to the 31st percentile of the distribution. Effects of slightly smaller magnitude can be computed for lecturing clarity. Consistently with the findings of other studies (Carrell and West, 2010; Krautmann and Sander, 1999; Weinberg, Fleisher, and Hashimoto, 2009), when we use the contemporaneous effects (even columns) the estimated coefficients turn positive and highly significant for all items (but workload). In other words, the teachers of classes that are associated with higher grades in their own exam receive better evaluations from their students. The magnitudes of these effects are smaller than those estimated for our benchmark measures: one standard deviation change in the contemporaneous teacher effect increases the evaluation of overall teaching quality by 24% of a standard deviation and the evaluation of lecturing clarity by 11%.

The results in Table 10 clearly challenge the validity of students' evaluations of professors as a measure of teaching quality. Even abstracting from the possibility that professors strategi-

Table 10: Teacher effectiveness and students' evaluations

	Teaching quality		Lecturing clarity		Teacher ability in generating interest		Course logistics		Course workload	
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Teacher effectiveness										
Benchmark	-0.496**	-	-0.249**	-	-0.552**	-	-0.124	-	-0.090	-
	(0.236)		(0.113)		(0.226)		(0.095)		(0.104)	
Contemporaneous	-	0.238***	-	0.116***	-	0.214***	-	0.078***	-	-0.007
		(0.055)		(0.029)		(0.044)		(0.019)		(0.025)
Class characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Classroom characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Teacher's characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Degree program dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Subject area dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Term dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Partial R2	0.019	0.078	0.020	0.075	0.037	0.098	0.013	0.087	0.006	0.001
Observations	230	230	230	230	230	230	230	230	230	230

Weighted OLS estimates. Observations are weighted by the number of collected questionnaires in each class. Bootstrapped standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

cally adjust their grades to please the students (a practice that is made difficult by the timing of the evaluations, that are always collected before the exam takes place, and by the fact that the evaluations are communicated to the teachers with a certain delay), it might still be possible that professors who make the classroom experience more enjoyable do that at the expense of true learning or fail to encourage students to exert effort. Alternatively, students might reward teachers who prepare them for the exam, that is teachers who teach to the test, even if this is done at the expenses of true learning. This interpretation is consistent with the results in Weinberg, Fleisher, and Hashimoto (2009), who provide evidence that students are generally unaware of the value of the material they have learned in a course, and it is the interpretation that we adopt to develop the theoretical framework of Section 6.

Of course, one may also argue that students' satisfaction is important *per se* and, even, that universities should aim at maximizing satisfaction rather than learning. The solution to the principal-agent problem obviously depends on the objective function the principal wants to maximize. A public university should in principle incorporate preferences of the society as a whole: in this case, we doubt that any social planner would prefer to increase satisfaction rather than promoting "true learning" and increasing human capital. In the case of a private university (like Bocconi), the objectives of the principal are not so obvious, but one could always think of shifting the principal-agent problem simply one level up, with the relevant relationship being the one between the policymaker and the private educational institutions he regulates. In this case, the objective of the policymaker would be to make sure that private institutions (especially if they receive public funds) promote learning and the accumulation of human capital rather than simply making their students happy.

5 Robustness checks

In this section we present robustness checks for our main results in Sections 3 and 4.

First, we investigate the role of students' dropout in the estimation of our measures of teacher effectiveness. In our main empirical analysis students who do not have a complete academic record are excluded. These are students who either dropped out of higher education

or have transfered to another university or are still working towards the completion of their programs, whose formal duration was 4 years. They total about 10% of all the students who enrolled in their first year in 1998-1999. In order to check that excluding them does not affect our main results, in Figure 2 we compare our benchmark measure of teacher effectiveness estimated in Section 3 with similar estimates that include such dropout students. As it is evident, the two sets of estimates are very similar and regressing one over the other (controlling for degree program, term and subject fixed effects) yields an R^2 of over 88%. Importantly, there does not seem to be larger discrepancies between the two versions of the class effects for the best or the worst teachers.

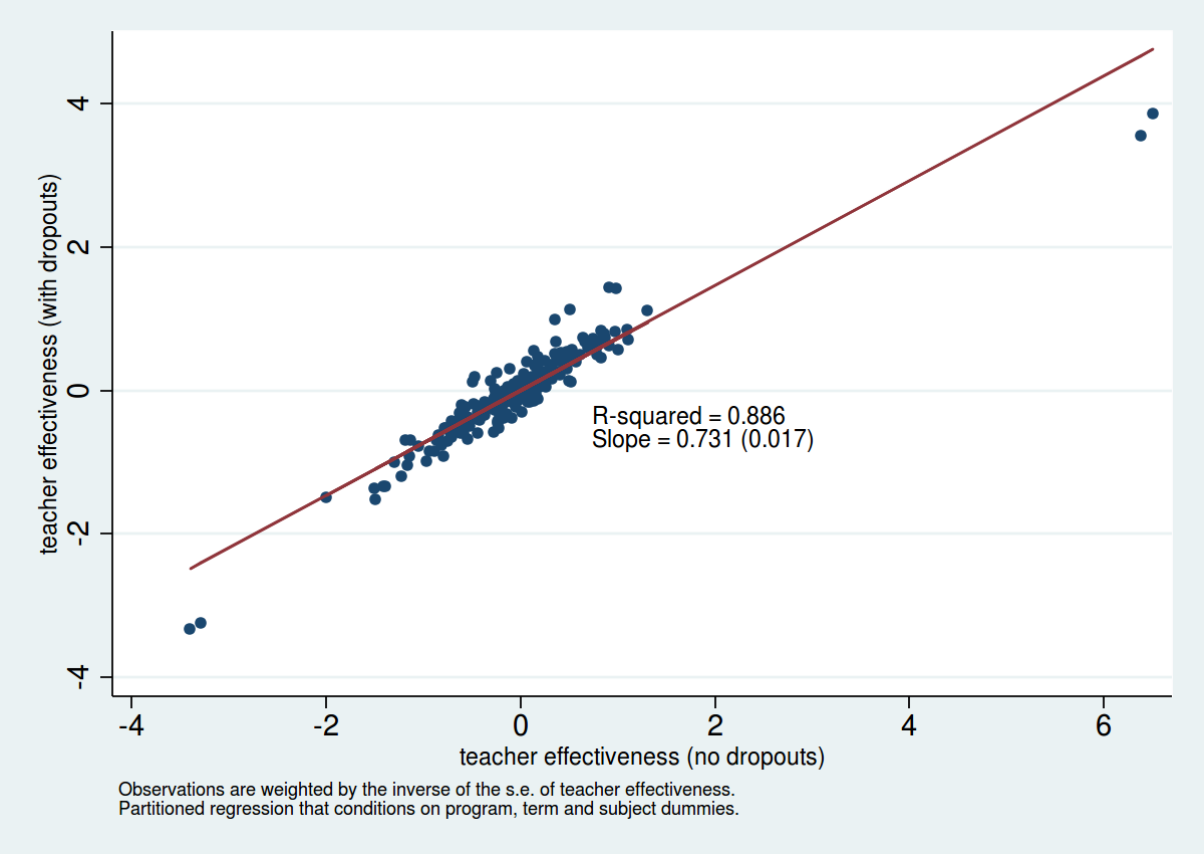


Figure 2: Robustness check for dropouts

Second, one might be worried that students might not comply with the random assignment to the classes. For various reasons they may decide to attend one or more courses in a different class from the one to which they were formally allocated. For example, they may desire to stay with their friends, who might have been assigned to a different class, or they may like a

specific teacher, who is known to present the subject particularly clearly. Unfortunately, such changes would not be recorded in our data, unless the student formally asked to be allocated to a different class, a request that needed to be adequately motivated.[29] Hence, we cannot exclude a priori that some students switch classes.

If the process of class switching is unrelated to teaching quality, then it merely affects the precision of our estimated class effects, but it is very well possible that students switch in search for good or lenient lecturers. We can get some indication of the extent of this problem from the students' answers to an item of the evaluation questionnaires that asks about the congestion in the classroom. Specifically, the question asks whether the number of students in the class was detrimental to one's learning. We can, thus, identify the most congested classes from the average answer to such question in each course.

Courses in which students concentrate in the class of one or few professors should be characterized by a very skewed distribution of such a measure of congestion, with one or a few classes being very congested and the others being pretty empty. Thus, for each course we compute the difference in the congestion indicator between the most and the least congested classes (over the standard deviation). Courses in which such a difference is very large should be the ones that are more affected by switching behaviors.

In Table 11 we replicate our benchmark estimates for the two core evaluation items (overall teaching quality and lecturing clarity) by excluding the most switched course (Panel B), i.e. the course with the largest difference between the most and the least congested classes (which is marketing). For comparison we also report the original estimates from Table 10 in Panel A and we find that results change only marginally. Next, in Panel C and D we exclude from the sample also the second most switched course (human resource management) and the five most switched courses, respectively.[30] Again, the estimated coefficients are only mildly affected, although the significance levels are reduced according with the smaller sample sizes. Overall, this exercise suggests that course switching should not affect our estimates in any major direction.

[29]Possible motivations for such requests could be health reasons. For example, due to a broken leg a student might not be able to reach classrooms in the upper floors of the university buildings and could ask to be assigned to a class taught on the ground floor.

[30]The five most switched courses are marketing, human resource management, mathematics for Economics and Management, financial mathematics and managerial accounting.

Table 11: Robustness check for class switching

	Overall teaching quality		Lecturing clarity	
	[1]	[2]	[3]	[4]
PANEL A: All courses				
Benchmark teacher effects	-0.496**	-	-0.249**	-
	(0.236)		(0.113)	
Contemporaneous teacher effects	-	0.238***	-	0.116***
		(0.055)		(0.029)
Observations	230	230	230	230
PANEL B: Excluding most switched course				
Benchmark teacher effects	-0.572**	-	-0.261**	-
	(0.267)		(0.118)	
Contemporaneous teacher effects	-	0.258***	-	0.121***
		(0.064)		(0.030)
Observations	222	222	222	222
PANEL C: Excluding most and second most switched course				
Benchmark teacher effects	-0.505*	-	-0.234*	-
	(0.272)		(0.128)	
Contemporaneous teacher effects	-	0.233***	-	0.112***
		(0.062)		(0.031)
Observations	214	214	214	214
PANEL D: Excluding five most switched courses				
Benchmark teacher effects	-0.579**	-	-0.229*	-
	(0.273)		(0.122)	
Contemporaneous teacher effects	-	0.154**	-	0.065**
		(0.063)		(0.032)
Observations	176	176	176	176

Weighted OLS estimates. Observations are weighted by the squared root of the number of collected questionnaires in each class.
Additional regressors: teacher characteristics (gender and coordinator status), class characteristics (class size, attendance, average high school grade, average entry test score, share of high ability students, share of students from outside Milan, share of top-income students), degree program dummies, term dummies, subject area dummies.
Bootstrapped standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Finally, one might be worried that our results may be generated by some endogenous reaction of students to the quality of their past teachers. For example, as one meets a bad teacher in one course one might be induced to exert higher effort in the future to catch up, especially if bad teaching resulted in a lower (contemporaneous) grade. Hence, the students evaluations may reflect real teaching quality and our measure of teacher effectiveness would be biased by such a process of mean reversion, leading to a negative correlation with real teaching quality and, consequently, also with the evaluations of the students.

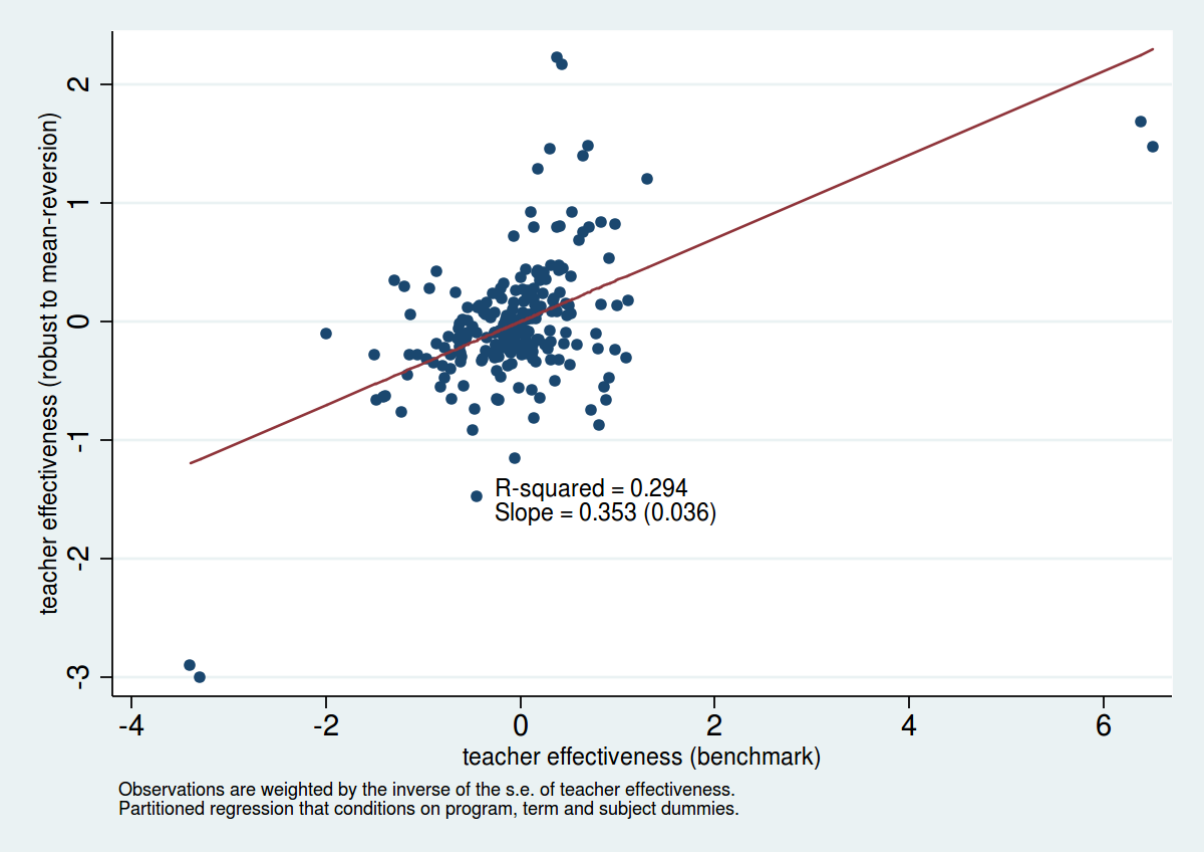


Figure 3: Robustness check for mean reversion in grades

To control for this potential feedback effect on students' effort, we recompute our benchmark measures of teacher effectiveness adding the student average grade in all previous courses to the set of controls. Figure 3 compares our benchmark teacher effectiveness with this augmented version, conditioning on the usual fixed effects for degree program, term and subject area and shows that the two are strongly correlated (even accounting for the outliers).

6 Interpreting the results: a simple theoretical framework

We think of teaching as the combination of two types of activities: *real teaching* and *teaching-to-the-test*. The first consists of presentations and discussions of the course material and leads to actual learning, conditional on the students exerting effort; the latter is aimed at maximizing performance in the exam, it requires lower effort by the students and it is not necessarily related to actual learning.

Practically, we think of real teaching as competent presentations of the course material with the aim of making students understand and master it and of teaching-to-the-test as mere repetition of exam tests and exercises with the aim of making students learn how to solve them, even without fully understanding their meaning.

Consider a setting in which teachers are heterogenous in their preference (or ability) to do real teaching. We measure such heterogeneity with a parameter $\mu_j \in [0, 1]$, such that a teacher j with $\mu_j = 0$ exclusively teaches to the test and a teacher with $\mu_j = 1$ exclusively engages in real teaching.

The grade x_i of a generic student i in the course taught by teacher (or in class) j is defined by the following production function:

$$x_i = \mu_j h(e_i) + (1 - \mu_j)\overline{x} \tag{4}$$

which is a linear combination of a function $h(\cdot)$ of student's effort e_i and a constant $\overline{x}$, weighted by the teacher's type μ_j. We assume $h(\cdot)$ to be a continuous and twice differentiable concave function. Under full real teaching ($\mu_j = 1$) grades vary with students' effort; on the other hand, if the teacher exclusively teaches to the test ($\mu_j = 0$), everyone gets the same grade $\overline{x}$, regardless of effort. This strong assumption can obviously be relaxed and all our implications will be maintained as long as the gradient of grades to effort increases with μ_j.

The parameter $\overline{x}$ measures the extent to which the exam material and the exam format lend themselves to teaching-to-the-test. To the one extreme, one can think of the exam as a selection of multiple-choice questions randomly drawn from a large pool. In such a situation, teaching-to-the-test merely consists in going over all the possible questions and memorizing the correct

answer. This is a setting which would feature a large $\overline{x}$. The other extreme are essays, where there is no obvious correct answers and one needs to personally and originally elaborate on one's own understanding of the course material. Of course, there are costs and benefits to each type of exam and multiple-choice tests are often adopted because they can be marked quickly, easily and uncontroversially. For the sake of simplicity, however, we abstract from cost-benefit considerations.

Furthermore, equation 4 assumes that teaching-to-the-test does not require students to exert effort. All our results would be qualitatively unchanged under the weaker assumption that teaching-to-the-test requires less effort by the students. We also assume that μ_j is a fixed characteristic of teacher j, so that the model effectively describes the conditions for selecting teachers of different types, a key piece of information for hiring and promotion decisions. Alternatively, μ_j could be treated as an endogenous variable under the control of the individual teacher, in which case the model would feature a rather standard agency problem where the university tries to provide incentives to the teachers to choose a μ_j close to 1. Although, such a model would be considerably more complicated than what we present in this section, its qualitative results would be unchanged and the limited information on teachers in our data would make its additional empirical content redundant in our setting.

More specifically, one could model μ_j as an endogenous choice of the teacher and generate heterogeneity by assuming that different activities (real teaching or teaching-to-the-test) require different efforts from the professors, who face heterogeneous marginal disutilities. Such an alternative model would feature both adverse selection and moral hazard and proper measurement of teaching quality could help addressing both issues, by facilitating the identification of low quality agents (high disutility of effort) and by incentivizing effort. In our simplified framework, only adverse selection of professors takes place, but the general intuition holds also in a more complicated setting.

In all cases, a key assumption is that μ_j is unobservable by the university administrators (the principal) and, although it might be observable to the students, it cannot be credibly communicated to third parties.

Assume now that students care about their grades but dislike exerting effort, so that the

utility function of a generic student i can be written as follows:

$$U_i = x_i - \frac{1}{2}\frac{e_i^2}{\eta_i} \tag{5}$$

where η_i is a measure of student's ability.

For simplicity, we assume that students are perfectly informed about the production function of grades, i.e. they know the type of their teacher, they know the return to their effort and there is no additional stochastic component to equation 4. This assumption can be easily relaxed by introducing either imperfect information about the teacher's type or about the exact specification of the production function and, consequently, by rewriting the utility function in equation 5 in expected terms. The main intuition of our results would be unchanged. Although the perfect information assumption is obviously a modeling device and does not correspond to reality, we do believe that students know a lot about their professors, either through conversations with older students or by observation through the duration of the course.

The utility function in equation 5 implicitly assumes that students are myopic, in the sense that they care only about grades and not about real learning. The main implications of the simple theory in this section would remain unchanged also with a different utility function that incorporates real learning, as long as students of different abilities care equally about it (just like they are equally myopic in the current specification).

The quasi-linearity of equation 5 simplifies the algebra of the model. Alternatively, we could have introduced some curvature in the utility function and assumed a linear production process without affecting the results. With non-linearities both in the production and in the utility functions one would have to make explicit a number of additional assumptions to guarantee existence and uniqueness of the equilibrium.

Students choose their optimal level of effort e_i^* according to the following first order conditions:

$$\mu_j \frac{\partial h(e)}{\partial e_i}(e_i^*) = \frac{e_i^*}{\eta_i} \tag{6}$$

Using equation 6 it is easy to derive the following results:

$$\frac{de_i^*}{d\eta_i} > 0 \tag{7}$$

$$\frac{de_i^*}{d\mu_j} > 0 \tag{8}$$

$$\frac{de_i^*}{d\mu_j d\eta_i} > 0 \tag{9}$$

Equation 7 shows that more able students exert higher effort. Equation 8 shows that more real teaching induces higher effort from the students and equation 9 indicates that such an effect is larger for the more able students

Additionally, using the envelope theorem it is easy to show that:

$$\frac{\partial U_i(e_i^*)}{\partial \mu_j} = h(e_i^*) - \overline{x} \tag{10}$$

Define $\overline{e}$ the level of effort such that $h(\overline{e}) = \overline{x}$. Moreover, since for a given μ_j there is a unique correspondence between effort and ability, $\overline{e}$ uniquely identifies a $\overline{\eta}$. Hence:

$$\frac{\partial U_i(e_i^*)}{\partial \mu_j} > 0 \quad \text{if} \quad \eta_i > \overline{\eta} \tag{11}$$

$$\frac{\partial U_i(e_i^*)}{\partial \mu_j} < 0 \quad \text{if} \quad \eta_i < \overline{\eta} \tag{12}$$

Equations 11 and 12 are particularly important under the assumption that, especially when answering questions about the overall quality of a course, students give a better evaluation to teachers (or classes) that are associated with a higher level of utility. Equations 11 and 12 suggest that high ability students evaluate better teachers or classes that are more focused on real learning while low ability students prefer teachers that teach to the test. Hence, if the (benchmark) teacher effects estimated in Section 3 indeed measure the real learning value of a class (μ_j, in the terminology of our model), we expect to see a more positive (or less negative) correlation between such class effects and the students' evaluations in those classes where the concentration of high ability students is higher.

7 Further evidence

In this section we present some additional pieces of evidence that are consistent with the implications of the model of Section 6.

First, in the model we assume that students evaluate professors on the basis of their realized utility from attending their courses. This might be a questionable assumption. Especially university administrators who organize and elaborate the students' questionnaires are often convinced that, when asked about the ability of the teacher in presenting the course material, students express their opinion regardless of whether the teacher has imposed a high effort cost on them in order to pass the exam. In fact, an alternative behavioral model would be one in which students observe the true type of the teacher and they truthfully communicate it in the questionnaires regardless of their individual classroom experience.

In order to provide support for our specification, in Table 12 we produce evidence that the students' evaluations respond to the weather conditions on the day when they were filled. There is ample evidence that people's utility (or welfare, happiness, satisfaction) improves with good meteorological conditions (Barrington-Leigh, 2008; Denissen, Butalid, Penke, and van Aken, 2008; Keller, Fredrickson, Ybarra, Coté, Johnson, Mikels, Conway, and Wager, 2005; Pray, 2011; Schwarz and Clore, 1983) and finding that such conditions also affect the evaluations of professors suggests that they indeed reflect utility rather than (or together with) teaching quality.

Specifically, we find that evaluations improve with temperature, deteriorate with rain and improve on foggy days. The effects are significant for most of the items that we consider and the signs of the estimates are consistent across items and specifications.

Obviously, teachers might be affected by meteorological conditions as much as their students and one may wonder whether the estimated effects in the odd columns of Table 12 reflect the indirect effect of the weather on teaching effectiveness. We consider this interpretation to be very unlikely since the questionnaires are distributed and filled before the lecture so that students should not be able to incorporate in their answers the performance of the teacher in the day the evaluation forms are filled in. Moreover, students' are asked to evaluate teachers'

Table 12: Students' evaluations and weather conditions

	Teaching quality		Lecturing clarity		Teacher ability in generating interest		Course logistics		Course workload	
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Av. temperature	0.139*	0.120	0.063*	0.054	0.171***	0.146***	0.051**	0.047**	0.057*	0.053*
	(0.074)	(0.084)	(0.036)	(0.038)	(0.059)	(0.054)	(0.020)	(0.019)	(0.031)	(0.029)
1=rain	-0.882**	-0.929**	-0.293	-0.314	-0.653**	-0.716**	-0.338***	-0.348***	0.081	0.071
	(0.437)	(0.417)	(0.236)	(0.215)	(0.327)	(0.287)	(0.104)	(0.108)	(0.109)	(0.128)
1=fog	0.741**	0.687*	0.391**	0.367**	0.008	-0.063	0.303***	0.292***	-0.254***	-0.265***
	(0.373)	(0.377)	(0.191)	(0.170)	(0.251)	(0.247)	(0.085)	(0.090)	(0.095)	(0.096)
Teaching	-	-0.424*	-	-0.189	-	-0.566**	-	-0.090	-	-0.088
effectiveness		(0.244)		(0.120)		(0.223)		(0.088)		(0.093)
Class characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Classroom characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Teacher's characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Degree program dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Subject area dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Term dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	230	230	230	230	230	230	230	230	230	230

Weighted OLS estimates. Observations are weighted by the number of collected questionnaires in each class. Bootstrapped standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

performance over the entire duration of the course and not exclusively on the day of the test.

Nevertheless, in the even columns of Table 12, we also condition on our benchmark measure of teaching effectiveness and, as we expected, we find that the estimated effects of both the weather conditions and teacher effectiveness itself change only marginally.

Second, our specification of the production function for exam grades in equation 4 implies a positive relationship between grade dispersion and the professor's propensity to engage in real teaching (μ_j). In our empirical exercise our measures of teacher effectiveness can be interpreted as measures of the μ_j's in the terminology of the model. Hence, if grades were more dispersed in the classes of the worst teachers one would have to question our specification of equation 4.

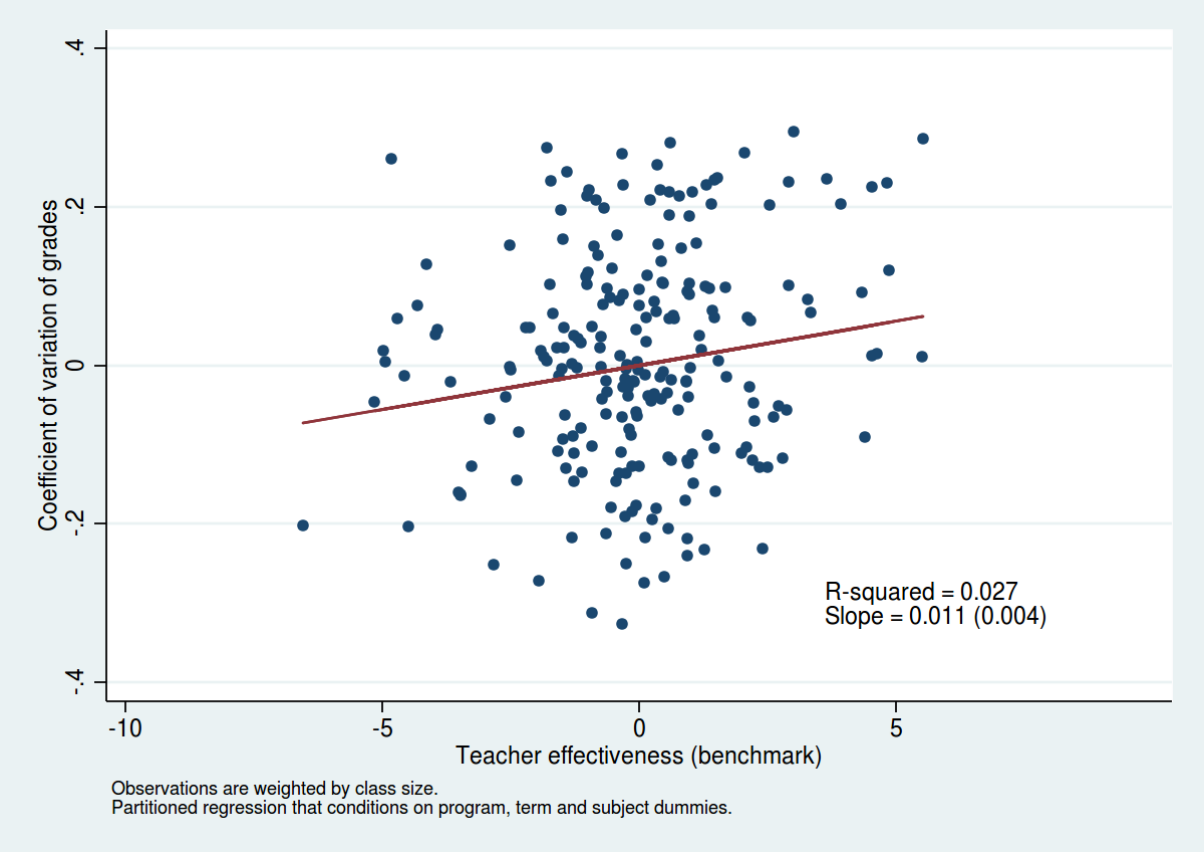


Figure 4: Teacher effectiveness and grade dispersion

In Figure 4 we plot the coefficient of variation of grades in each class (on the vertical axis) against our measure of teacher effectiveness (on the horizontal axis). To take proper account of differences across degree programs, the variables on both axes are the residuals of weighted OLS regressions that condition on degree program, term and subject area fixed effects, as in

standard partitioned regressions (the weights are the squared roots of class sizes). Consistently with equation 4 in our model, the two variables are positively correlated and such a correlation is statistically significant at conventional levels: a simple univariate OLS regression of the variable on the vertical axis on the variable on the horizontal axis yields a coefficient of 0.011 with a standard error of 0.004.

Table 13: Teacher effectiveness and students evaluations by share of high ability students

	Presence of high-ability students			
	all	>0.22 (top 75%)	>0.25 (top 50%)	>0.27 (top 25%)
	[1]	[2]	[3]	[4]
PANEL A: Overall teaching quality				
Teaching effectiveness	-0.496**	-0.502*	-0.543	-0.141***
	(0.236)	(0.310)	(0.439)	(0.000)
PANEL B: Lecturing clarity				
Teaching effectiveness	-0.249**	-0.240	-0.283	-0.116*
	(0.113)	(0.140)	(0.191)	(0.068)
Observations	230	171	114	56

Weighted OLS estimates. Observations are weighted by the squared root of the number of collected questionnaires in each class.
Additional regressors: teacher characteristics (gender and coordinator status), class characteristics (class size, attendance, average high school grade, average entry test score, share of high ability students, share of students from outside Milan, share of top-income students), degree program dummies, term dummies, subject area dummies.
Bootstrapped standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Next, according to equations 11 and 12, we expect the correlation between our measures of teacher effectiveness and the average student evaluations to be less negative in classes where the share of high ability students is higher. This is the hypothesis that we investigate in Table 13. We define as high ability those students who score in the upper quartile of the distribution of the entry test score and, for each class in our data, we compute the share of such students. Then, we investigate the relationship between the students' evaluations and teacher effectiveness by restricting the sample to classes in which high-ability students are over-represented. Results seem to suggest the presence of non linearities or threshold effects, as the estimated coefficient remains relatively stable until the fraction of high ability students in the class goes above 27%.

At that point, the estimated effect of teacher effectiveness on students' evaluations is about a quarter of the one estimated on the entire sample. The results, thus, suggest that the negative correlations reported in Table 10 are mostly due to classes with a particularly low incidence of high ability students.

8 Conclusions

Using administrative archives from Bocconi University and exploiting random variation in students' allocation to teachers within courses we find that, on average, students evaluate positively classes that give high grades and negatively classes that are associated with high grades in subsequent courses. These empirical findings can be rationalized with a simple model featuring heterogeneity in the preferences (or ability) of teachers to engage in real teaching rather than teaching-to-the-test, with the former requiring higher effort from students than the latter. Furthermore, we also find that weather conditions on the day the questionnaires are filled in are correlated with students' evaluations of theachers. This is consistent with the assumption of our model, namely that students' evaluations reflect students' perceived utility more than teachers' ability. Overall, our results cast serious doubts on the validity of students' evaluations of professors as measures of teaching quality or effort.

At the same time, the strong effects of teaching quality on students' outcomes, as documented in Section 3, suggest that improving the quantity or the quality of professors' inputs in the education production function can lead to large gains. Under the interpretation offered by our model in Section 6, this could be achieved through various types of interventions. For example, one may think of adopting exam formats that reduce the returns to teaching-to-the-test, although this may come at larger costs due to the additional time needed to grade less standardized tests.

Alternatively, one may stick to the use of students' evaluations to measure teachers' performance but limit the extent of grade leniency that may be induced in such a system, for example by making sure that teaching and grading are done by different persons. Anecdotically, we know that at Bocconi it is common practice among the teachers of the core statistics course to

randomize the grading, i.e. at the end of the course the teachers of the different classes are randomly assigned the papers of another class for marking. In the only year in which this practice was abandoned, average grades increased substantially.

Another variation to the current most common use of the students' evaluations consists in postponing the collection of students' opinions, so as to give them time to appreciate the value of real teaching in subsequent learning (or even in the market). Obviously, this would also pose problems in terms of recall bias and possible retaliation for low grading.

Alternatively, one may also think of other forms of performance measurement that are more in line with the peer-review approach adopted in the evaluation of research output. It is already common practice in several departments to have colleagues sitting in some classes and observing teacher performance, especially of assistant professors. This is often done primarily with the aim of offering advise, but in principle it could also be used to measure outcomes. An obvious concern is that one could change behavior due to the presence of the observer. A slightly more sophisticated version of the same method could be based on the use of cameras to record a few teaching sessions during the course without the teacher knowing exactly which ones. The video recordings could then be viewed and evaluated by an external professor in the same field.

Obviously, these, as well as other potential alternative measurement methods, are costly but they should be compared with the costs of the current systems of collecting students' opinions about teachers, which are often non trivial.

References

BAKER, G., R. GIBBONS, AND K. J. MURPHY (1994): "Subjective performance measures in optimal incentive contracts," *Quarterly Journal of Economics*, 109(4), 1125–1156.

BARRINGTON-LEIGH, C. (2008): "Weather as a transient influence on survey-reported satisfaction with life," Draft research paper, University of British Columbia.

BECKER, W. E., AND M. WATTS (1999): "How departments of economics should evaluate teaching," *American Economic Review (Papers and Proceedings)*, 89(2), 344–349.

BROWN, B. W., AND D. H. SAKS (1987): "The microeconomics of the allocation of teachers' time and student learning," *Economics of Education Review*, 6(4), 319–332.

CARRELL, S. E., AND J. E. WEST (2010): "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," *Journal of Political Economy*, 118(3), 409–32.

CHETTY, R., J. N. FRIEDMAN, N. HILGER, E. SAEZ, D. W. SCHANZENBACH, AND D. YAGAN (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence From Project STAR," *Quarterly Journal of Economics*, forthcoming.

DE GIORGI, G., M. PELLIZZARI, AND S. REDAELLI (2010): "Identification of Social Interactions through Partially Overlapping Peer Groups," *American Economic Journal: Applied Economics*, 2(2), 241–275.

DE GIORGI, G., M. PELLIZZARI, AND W. G. WOOLSTON (2011): "Class Size and Class Heterogeneity," *Journal of the European Economic Association*, forthcoming.

DENISSEN, J. J. A., L. BUTALID, L. PENKE, AND M. A. VAN AKEN (2008): "The Effects of Weather on Daily Mood: A Multilevel Approach," *Emotion*, 8, 662–667.

DUFLO, E., R. HANNA, AND M. KREMER (2010): "Incentives Work: Getting Teachers to Come to School," mimeo, MIT.

FIGLIO, D. N., AND L. KENNY (2007): "Individual teacher incentives and student performance," *Journal of Public Economics*, 91, 901–914.

GOLDHABER, D., AND M. HANSEN (2010): "Using performance on the job to inform teacher tenure decisions," *American Economic Review (Papers and Proceedings)*, 100(2), 250–255.

HANUSHEK, E. A. (1979): "Conceptual and empirical issues in the estimation of educational production functions," *Journal of Human Resources*, 14, 351–388.

HANUSHEK, E. A., AND S. G. RIVKIN (2006): "Teacher quality," in *Handbook of the Economics of Education*, ed. by E. A. Hanushek, and F. Welch, vol. 1, pp. 1050–1078. North Holland, Amsterdam.

——— (2010): "Generalizations about using value-added measures of teacher quality," *American Economic Review (Papers and Proceedings)*, 100(2), 267–271.

HIRSCH, J. E. (2005): "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569–16572.

HOFFMAN, F., AND P. OREOPOULOS (2009): "Professor Qualities and Student Achievement," *The Review of Economics and Statistics*, 91(1), 83–92.

HOGAN, T. D. (1981): "Faculty research activity and the quality of graduate training," *Journal of Human Resources*, 16(3), 400–415.

HOLMSTROM, B., AND P. MILGROM (1994): "The firm as an incentive system," *American Economic Review*, 84(4), 972–991.

JACOB, B. A., AND L. LEFGREN (2008): "Can principals identify effective teachers? Evidence on subjective performance evaluation in education," *Journal of Labor Economics*, 26, 101–136.

KANE, T. J., AND D. O. STAIGER (2008): "Estimating teacher impacts on student achievement: an experimental evaluation," Discussion Paper 14607, NBER Working Paper Series.

KELLER, M. C., B. L. FREDRICKSON, O. YBARRA, S. COTÉ, K. JOHNSON, J. MIKELS, A. CONWAY, AND T. WAGER (2005): "A Warm Heart and a Clear Head. The Contingent Effects of Weather on Mood and Cognition," *Psychological Science*, 16(9), 724–731.

KRAUTMANN, A. C., AND W. SANDER (1999): "Grades and student evaluations of teachers," *Economics of Education Review*, 18, 59–63.

KRUEGER, A. B. (1999): "Experimental estimates of education production functions," *Quarterly Journal of Economics*, 114, 497–532.

LAVY, V. (2009): "Performance Pay and Teachers' Effort, Productivity and Grading Ethics," *The American Economic Review*, 95(5), 1979–2011.

MULLIS, I. V., M. O. MARTIN, D. F. ROBITAILLE, AND P. FOY (2009): *TIMSS Advanced 2008 International Report*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

OECD (2008): *Education at a Glance*. Organization of Economic Cooperation and Development, Paris.

——— (2010): *PISA 2009 at a Glance*. OECD Publishing.

PRAY, M. C. (2011): "Some Like It Mild and Not Too Wet: the Influence of Weather on Subjective Well-Being," Cahiers de recherche 1116, CIRPEE.

PRENDERGAST, C., AND R. H. TOPEL (1996): "Favoritism in organizations," *Journal of Political Economy*, 104(5), 958–978.

RIVKIN, S. G., E. A. HANUSHEK, AND J. F. KAIN (2005): "Teachers, Schools and Acedemic Achievement," *Econometrica*, 73(2), 417–458.

ROCKOFF, J. E. (2004): "The impact of individual teachers on student achievement: evidence from panel data," *American Economic Review (Papers and Proceedings)*, 94(2), 247–252.

ROCKOFF, J. E., AND C. SPERONI (2010): "Subjective and Objective Evaluations of Teacher Effectiveness," *American Economic Review (Papers and Proceedings)*, 100(2), 261–266.

ROTHSTEIN, J. (2009): "Student sorting and bias in value added estimation: selection on observables and unobservables," *Education Finance and Policy*, 4(4), 537–571.

——— (2010): "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics*, 125(1), 175–214.

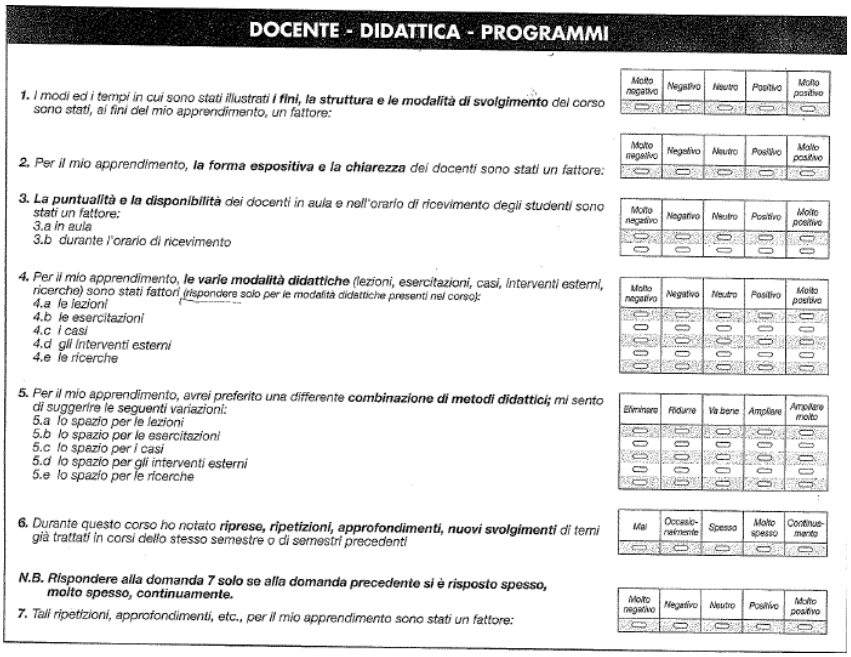
SCHWARZ, N., AND G. L. CLORE (1983): "Mood, Misattribution, and Judgments of Well-being: Informative and Directive Functions of Affective States," *Journal of Personality and Social Psychology*, 45(3), 513–523.

SWAMY, P. A. V. B., AND S. S. ARORA (1972): "The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models," *Econometrica*, 40(2), pp. 261–275.

TYLER, J. H., E. S. TAYLOR, T. J. KANE, AND A. L. WOOTEN (2010): "Using student performance data to identify effective classroom practices," *American Economic Review (Papers and Proceedings)*, 100(2), 256–260.

WEINBERG, B. A., B. M. FLEISHER, AND M. HASHIMOTO (2009): "Evaluating Teaching in Higher Education," *Journal of Economic Education*, 40(3), 227–261.

Appendix



DOCENTE - DIDATTICA - PROGRAMMI

1. *I modi ed i tempi in cui sono stati illustrati* ***i fini, la struttura e le modalità di svolgimento*** *del corso sono stati, ai fini del mio apprendimento, un fattore:*
   Molto negativo | Negativo | Neutro | Positivo | Molto positivo

2. *Per il mio apprendimento,* ***la forma espositiva e la chiarezza*** *dei docenti sono stati un fattore:*
   Molto negativo | Negativo | Neutro | Positivo | Molto positivo

3. ***La puntualità e la disponibilità*** *dei docenti in aula e nell'orario di ricevimento degli studenti sono stati un fattore:*
   3.a in aula
   3.b durante l'orario di ricevimento
   Molto negativo | Negativo | Neutro | Positivo | Molto positivo

4. *Per il mio apprendimento,* ***le varie modalità didattiche*** *(lezioni, esercitazioni, casi, interventi esterni, ricerche) sono stati fattori (rispondere solo per le modalità didattiche presenti nel corso):*
   4.a le lezioni
   4.b le esercitazioni
   4.c i casi
   4.d gli interventi esterni
   4.e le ricerche
   Molto negativo | Negativo | Neutro | Positivo | Molto positivo

5. *Per il mio apprendimento, avrei preferito una differente* ***combinazione di metodi didattici;*** *mi sento di suggerire le seguenti variazioni:*
   5.a lo spazio per le lezioni
   5.b lo spazio per le esercitazioni
   5.c lo spazio per i casi
   5.d lo spazio per gli interventi esterni
   5.e lo spazio per le ricerche
   Eliminare | Ridurre | Va bene | Ampliare | Ampliare molto

6. *Durante questo corso ho notato* ***riprese, ripetizioni, approfondimenti, nuovi svolgimenti*** *di temi già trattati in corsi dello stesso semestre o di semestri precedenti*
   Mai | Occasionalmente | Spesso | Molto spesso | Continuamente

N.B. Rispondere alla domanda 7 solo se alla domanda precedente si è risposto spesso, molto spesso, continuamente.

7. *Tali ripetizioni, approfondimenti, etc., per il mio apprendimento sono stati un fattore:*
   Molto negativo | Negativo | Neutro | Positivo | Molto positivo

Figure A-1: Excerpt of student questionnaire

Table A-1: Descriptive statistics of estimated class effects

	Management	Economics	Law & Management	Total
Std. dev. of estimated class effects				
mean	0.054	0.157	0.035	0.081
minimum	0.029	0.058	0.004	0.004
maximum	0.092	0.241	0.087	0.241
Largest minus smallest class effect				
mean	0.152	0.423	0.050	0.211
minimum	0.045	0.010	0.005	0.005
maximum	0.249	0.723	0.122	0.723
No. of courses	20	11	7	38
No. of classes	144	72	14	230

Table A-2: Descriptive statistics of *subject* teacher effectiveness

	Management	Economics	Law & Management	Total
PANEL A: Std. dev. of estimated teacher effects				
mean	0.095	0.244	0.099	0.140
minimum	0.055	0.049	0.018	0.018
maximum	0.163	0.342	0.194	0.342
PANEL B: Largest minus smallest teacher effect				
mean	0.266	0.733	0.140	0.377
minimum	0.175	0.069	0.026	0.026
maximum	0.428	1.171	0.275	1.171
No. of courses	17	10	7	34
No. of classes	128	70	14	212

Table A-3: Descriptive statistics of *contemporaneous* teacher effectiveness

	Management	Economics	Law & Management	Total
PANEL A: Std. dev. of estimated teacher effects				
mean	0.200	0.310	0.163	0.225
minimum	0.094	0.150	0.001	0.001
maximum	0.351	0.507	0.468	0.507
PANEL B: Largest minus smallest teacher effect				
mean	0.553	0.819	0.231	0.571
minimum	0.133	0.213	0.001	0.001
maximum	1.041	1.626	0.661	1.626
No. of courses	20	11	7	38
No. of classes	144	72	14	230

Table A-4: Wording of the evaluation questions

Overall teaching quality	*On a scale 0 to 10, provide your overall evaluation of the course you attended in terms of quality of the teaching.*
Clarity of the lectures	*On a scale 1 to 5, where 1 means complete disagreement and 5 complete agreement, indicate to what extent you agree with the following statement: the speech and the language of the teacher during the lectures are clear and easily understandable.*
Ability in generating interest for the subject	*On a scale 0 to 10, provide your overall evaluation about the teacher's ability in generating interest for the subject*
Logistics of the course	*On a scale 1 to 5, where 1 means complete disagreement and 5 complete agreement, indicate to what extent you agree with the following statement: the course has been carried out coherently with the objectives, the content and the schedule that were communicated to us at the beginning of the course by the teacher.*
Workload of the course	*On a scale 1 to 5, where 1 means complete disagreement and 5 complete agreement, indicate to what extent you agree with the following statement: the amount of study materials required for the preparation of the exam has been realistically adequate to the objective of learning and sitting the exams of all courses of the term.*

Table A-5: Correlations between evaluations items

	Overall teaching quality	Lecturing clarity	Teacher generates interest	Course logistics	Course workload
Overall teaching quality	1.000	-	-	-	-
Lecturing clarity	0.888 (0.000)	1.000	-	-	-
Teacher generates interest	0.697 (0.000)	0.536 (0.000)	1.000	-	-
Course logistics	0.742 (0.000)	0.698 (0.000)	0.506 (0.000)	1.000	-
Course workload	0.124 (0.060)	0.122 (0.064)	0.193 (0.003)	0.094 (0.153)	1.000

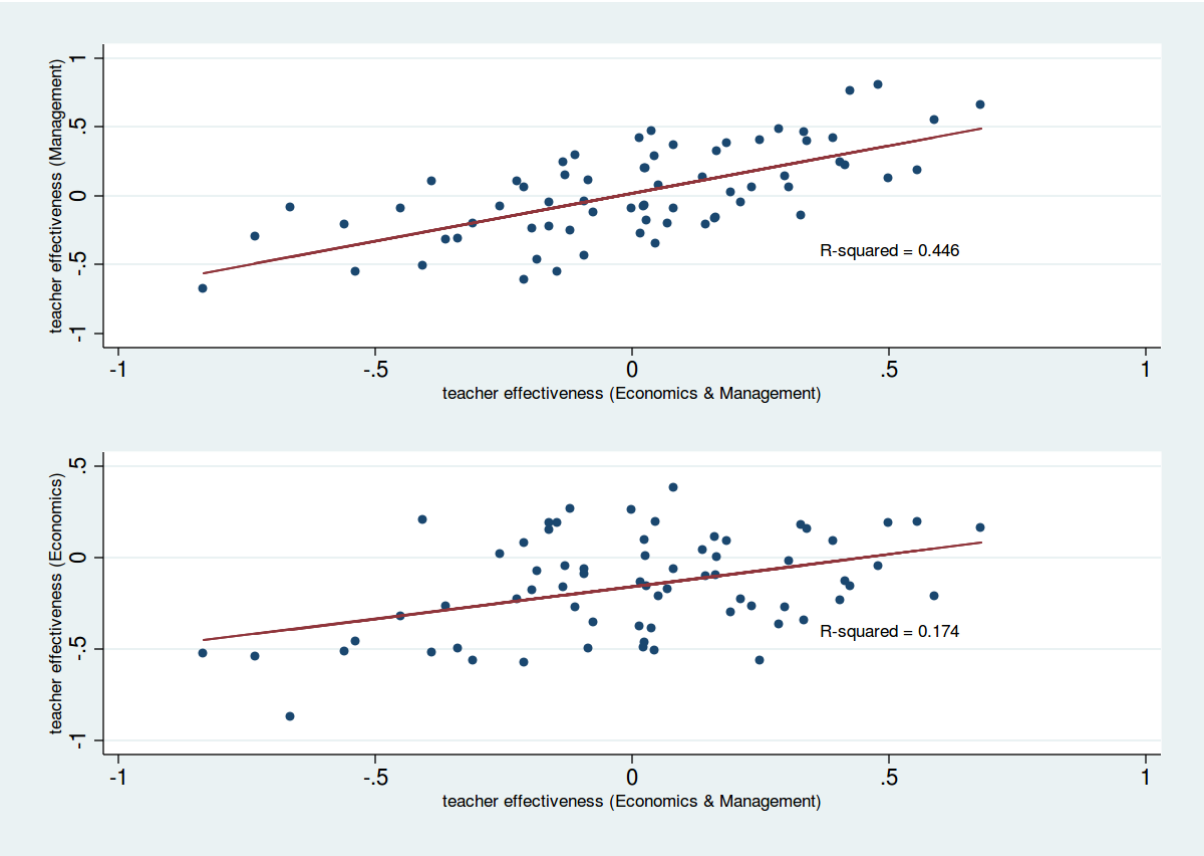


Figure A-2: Economics and Management common courses - Benchmark teacher effectiveness

RECENTLY PUBLISHED "TEMI" (*)

N. 804 – *Il miglioramento qualitativo delle produzioni italiane: evidenze da prezzi e strategie delle imprese*, by Valter di Giacinto and Giacinto Micucci (April 2011).

N. 805 – *What determines annuity demand at retirement?*, by Giuseppe Cappelletti, Giovanni Guazzarotti and Pietro Tommasino (April 2011).

N. 806 – *Heterogeneity and learning with complete markets*, by Sergio Santoro (April 2011).

N. 807 – *Housing, consumption and monetary policy: how different are the U.S. and the euro area?*, by Alberto Musso, Stefano Neri and Livio Stracca (April 2011).

N. 808 – *The monetary transmission mechanism in the euro area: has it changed and why?*, by Martina Cecioni and Stefano Neri (April 2011).

N. 809 – *Convergence clubs, the euro-area rank and the relationship between banking and real convergence*, by Massimiliano Affinito (June 2011).

N. 810 – *The welfare effect of foreign monetary conservatism with non-atomistic wage setters*, by Vincenzo Cuciniello (June 2011).

N. 811 – *Schooling and youth mortality: learning from a mass military exemption*, by Piero Cipollone and Alfonso Rosolia (June 2011).

N. 812 – *Welfare costs of inflation and the circulation of US currency abroad*, by Alessandro Calza and Andrea Zaghini (June 2011).

N. 813 – *Legal status of immigrants and criminal behavior: evidence from a natural experiment*, by Giovanni Mastrobuoni and Paolo Pinotti (June 2011).

N. 814 – *An unexpected crisis? Looking at pricing effectiveness of different banks*, by Valerio Vacca (July 2011).

N. 815 – *Skills or culture? An analysis of the decision to work by immigrant women in Italy*, by Antonio Accetturo and Luigi Infante (July 2011).

N. 816 – *Home bias in interbank lending and banks' resolution regimes*, by Michele Manna (July 2011).

N. 817 – *Macroeconomic determinants of carry trade activity*, by Alessio Anzuini and Fabio Fornari (September 2011).

N. 818 – *Leaving home and housing prices. The experience of Italian youth emancipation*, by Francesca Modena and Concetta Rondinelli (September 2011).

N. 819 – *The interbank market after the financial turmoil: squeezing liquidity in a "lemons market" or asking liquidity "on tap"*, by Antonio De Socio (September 2011).

N. 820 – *The relationship between the PMI and the Italian index of industrial production and the impact of the latest economic crisis*, by Valentina Aprigliano (September 2011).

N. 821 – *Inside the sovereign credit default swap market: price discovery, announcements, market behaviour and corporate sector*, by Alessandro Carboni (September 2011).

N. 822 – *The demand for energy of Italian households*, by Ivan Faiella (September 2011).

N. 823 – *Sull'ampiezza ottimale delle giurisdizioni locali: il caso delle province italiane*, by Guglielmo Barone (September 2011).

N. 824 – *The public-private pay gap: a robust quantile approach*, by Domenico Depalo and Raffaela Giordano (September 2011).

(*) Requests for copies should be sent to:
Banca d'Italia – Servizio Studi di struttura economica e finanziaria – Divisione Biblioteca e Archivio storico – Via Nazionale, 91 – 00184 Rome – (fax 0039 06 47922059). They are available on the Internet www.bancaditalia.it.

2008

P. ANGELINI, *Liquidity and announcement effects in the euro area*, Giornale degli Economisti e Annali di Economia, v. 67, 1, pp. 1-20, **TD No. 451 (October 2002).**

P. ANGELINI, P. DEL GIOVANE, S. SIVIERO and D. TERLIZZESE, *Monetary policy in a monetary union: What role for regional information?*, International Journal of Central Banking, v. 4, 3, pp. 1-28, **TD No. 457 (December 2002).**

F. SCHIVARDI and R. TORRINI, *Identifying the effects of firing restrictions through size-contingent Differences in regulation*, Labour Economics, v. 15, 3, pp. 482-511, **TD No. 504 (June 2004).**

L. GUISO and M. PAIELLA,, *Risk aversion, wealth and background risk*, Journal of the European Economic Association, v. 6, 6, pp. 1109-1150, **TD No. 483 (September 2003).**

C. BIANCOTTI, G. D'ALESSIO and A. NERI, *Measurement errors in the Bank of Italy's survey of household income and wealth*, Review of Income and Wealth, v. 54, 3, pp. 466-493, **TD No. 520 (October 2004).**

S. MOMIGLIANO, J. HENRY and P. HERNÁNDEZ DE COS, *The impact of government budget on prices: Evidence from macroeconometric models*, Journal of Policy Modelling, v. 30, 1, pp. 123-143 **TD No. 523 (October 2004).**

L. GAMBACORTA, *How do banks set interest rates?*, European Economic Review, v. 52, 5, pp. 792-819, **TD No. 542 (February 2005).**

P. ANGELINI and A. GENERALE, *On the evolution of firm size distributions*, American Economic Review, v. 98, 1, pp. 426-438, **TD No. 549 (June 2005).**

R. FELICI and M. PAGNINI, *Distance, bank heterogeneity and entry in local banking markets*, The Journal of Industrial Economics, v. 56, 3, pp. 500-534, **No. 557 (June 2005).**

S. DI ADDARIO and E. PATACCHINI, *Wages and the city. Evidence from Italy*, Labour Economics, v.15, 5, pp. 1040-1061, **TD No. 570 (January 2006).**

S. SCALIA, *Is foreign exchange intervention effective?*, Journal of International Money and Finance, v. 27, 4, pp. 529-546, **TD No. 579 (February 2006).**

M. PERICOLI and M. TABOGA, *Canonical term-structure models with observable factors and the dynamics of bond risk premia*, Journal of Money, Credit and Banking, v. 40, 7, pp. 1471-88, **TD No. 580 (February 2006).**

E. VIVIANO, *Entry regulations and labour market outcomes. Evidence from the Italian retail trade sector*, Labour Economics, v. 15, 6, pp. 1200-1222, **TD No. 594 (May 2006).**

S. FEDERICO and G. A. MINERVA, *Outward FDI and local employment growth in Italy*, Review of World Economics, Journal of Money, Credit and Banking, v. 144, 2, pp. 295-324, **TD No. 613 (February 2007).**

F. BUSETTI and A. HARVEY, *Testing for trend*, Econometric Theory, v. 24, 1, pp. 72-87, **TD No. 614 (February 2007).**

V. CESTARI, P. DEL GIOVANE and C. ROSSI-ARNAUD, *Memory for prices and the Euro cash changeover: an analysis for cinema prices in Italy*, In P. Del Giovane e R. Sabbatini (eds.), The Euro Inflation and Consumers' Perceptions. Lessons from Italy, Berlin-Heidelberg, Springer, **TD No. 619 (February 2007).**

B. H. HALL, F. LOTTI and J. MAIRESSE, *Employment, innovation and productivity: evidence from Italian manufacturing microdata*, Industrial and Corporate Change, v. 17, 4, pp. 813-839, **TD No. 622 (April 2007).**

J. SOUSA and A. ZAGHINI, *Monetary policy shocks in the Euro Area and global liquidity spillovers*, International Journal of Finance and Economics, v.13, 3, pp. 205-218, **TD No. 629 (June 2007).**

M. DEL GATTO, GIANMARCO I. P. OTTAVIANO and M. PAGNINI, *Openness to trade and industry cost dispersion: Evidence from a panel of Italian firms*, Journal of Regional Science, v. 48, 1, pp. 97-129, **TD No. 635 (June 2007).**

P. DEL GIOVANE, S. FABIANI and R. SABBATINI, *What's behind "inflation perceptions"? A survey-based analysis of Italian consumers*, in P. Del Giovane e R. Sabbatini (eds.), *The Euro Inflation and Consumers' Perceptions. Lessons from Italy*, Berlin-Heidelberg, Springer, **TD No. 655 (January 2008).**

R. BRONZINI, G. DE BLASIO, G. PELLEGRINI and A. SCOGNAMIGLIO, *La valutazione del credito d'imposta per gli investimenti*, Rivista di politica economica, v. 98, 4, pp. 79-112, **TD No. 661 (April 2008).**

B. BORTOLOTTI, and P. PINOTTI, *Delayed privatization,* Public Choice, v. 136, 3-4, pp. 331-351, **TD No. 663 (April 2008).**

R. BONCI and F. COLUMBA, *Monetary policy effects: New evidence from the Italian flow of funds,* Applied Economics , v. 40, 21, pp. 2803-2818, **TD No. 678 (June 2008).**

M. CUCCULELLI, and G. MICUCCI, *Family Succession and firm performance: evidence from Italian family firms,* Journal of Corporate Finance, v. 14, 1, pp. 17-31, **TD No. 680 (June 2008).**

A. SILVESTRINI and D. VEREDAS, *Temporal aggregation of univariate and multivariate time series models: a survey,* Journal of Economic Surveys, v. 22, 3, pp. 458-497, **TD No. 685 (August 2008).**

2009

F. PANETTA, F. SCHIVARDI and M. SHUM, *Do mergers improve information? Evidence from the loan market,* Journal of Money, Credit, and Banking, v. 41, 4, pp. 673-709, **TD No. 521 (October 2004).**

M. BUGAMELLI and F. PATERNÒ, *Do workers' remittances reduce the probability of current account reversals?*, World Development, v. 37, 12, pp. 1821-1838, **TD No. 573 (January 2006).**

P. PAGANO and M. PISANI, *Risk-adjusted forecasts of oil prices*, The B.E. Journal of Macroeconomics, v. 9, 1, Article 24, **TD No. 585 (March 2006).**

M. PERICOLI and M. SBRACIA, *The CAPM and the risk appetite index: theoretical differences, empirical similarities, and implementation problems,* International Finance, v. 12, 2, pp. 123-150, **TD No. 586 (March 2006).**

U. ALBERTAZZI and L. GAMBACORTA, *Bank profitability and the business cycle,* Journal of Financial Stability, v. 5, 4, pp. 393-409, **TD No. 601 (September 2006).**

S. MAGRI, *The financing of small innovative firms: the Italian case*, Economics of Innovation and New Technology, v. 18, 2, pp. 181-204, **TD No. 640 (September 2007).**

V. DI GIACINTO and G. MICUCCI, *The producer service sector in Italy: long-term growth and its local determinants,* Spatial Economic Analysis, Vol. 4, No. 4, pp. 391-425, **TD No. 643 (September 2007).**

F. LORENZO, L. MONTEFORTE and L. SESSA, *The general equilibrium effects of fiscal policy: estimates for the euro area*, Journal of Public Economics, v. 93, 3-4, pp. 559-585, **TD No. 652 (November 2007).**

R. GOLINELLI and S. MOMIGLIANO, *The Cyclical Reaction of Fiscal Policies in the Euro Area. A Critical Survey of Empirical Research*, Fiscal Studies, v. 30, 1, pp. 39-72, **TD No. 654 (January 2008).**

P. DEL GIOVANE, S. FABIANI and R. SABBATINI, *What's behind "Inflation Perceptions"? A survey-based analysis of Italian consumers*, Giornale degli Economisti e Annali di Economia, v. 68, 1, pp. 25-52, **TD No. 655 (January 2008).**

F. MACCHERONI, M. MARINACCI, A. RUSTICHINI and M. TABOGA, *Portfolio selection with monotone mean-variance preferences,* Mathematical Finance, v. 19, 3, pp. 487-521, **TD No. 664 (April 2008).**

M. AFFINITO and M. PIAZZA, *What are borders made of? An analysis of barriers to European banking integration*, in P. Alessandrini, M. Fratianni and A. Zazzaro (eds.): The Changing Geography of Banking and Finance, Dordrecht Heidelberg London New York, Springer, **TD No. 666 (April 2008).**

A. BRANDOLINI, *On applying synthetic indices of multidimensional well-being: health and income inequalities in France, Germany, Italy, and the United Kingdom*, in R. Gotoh and P. Dumouchel (eds.), Against Injustice. The New Economics of Amartya Sen, Cambridge, Cambridge University Press, **TD No. 668 (April 2008).**

G. FERRERO and A. NOBILI, *Futures contract rates as monetary policy forecasts,* International Journal of Central Banking, v. 5, 2, pp. 109-145, **TD No. 681 (June 2008).**

P. CASADIO, M. LO CONTE and A. NERI, *Balancing work and family in Italy: the new mothers' employment decisions around childbearing*, in T. Addabbo and G. Solinas (eds.), Non-Standard Employment and Qualità of Work, Physica-Verlag. A Sprinter Company, **TD No. 684 (August 2008).**

L. ARCIERO, C. BIANCOTTI, L. D'AURIZIO and C. IMPENNA, *Exploring agent-based methods for the analysis of payment systems: A crisis model for StarLogo TNG,* Journal of Artificial Societies and Social Simulation, v. 12, 1, **TD No. 686 (August 2008).**

A. CALZA and A. ZAGHINI, *Nonlinearities in the dynamics of the euro area demand for M1,* Macroeconomic Dynamics, v. 13, 1, pp. 1-19, **TD No. 690 (September 2008).**

L. FRANCESCO and A. SECCHI, *Technological change and the households' demand for currency*, Journal of Monetary Economics, v. 56, 2, pp. 222-230, **TD No. 697 (December 2008).**

G. ASCARI and T. ROPELE, *Trend inflation, taylor principle, and indeterminacy*, Journal of Money, Credit and Banking, v. 41, 8, pp. 1557-1584, **TD No. 708 (May 2007).**

S. COLAROSSI and A. ZAGHINI, *Gradualism, transparency and the improved operational framework: a look at overnight volatility transmission*, International Finance, v. 12, 2, pp. 151-170, **TD No. 710 (May 2009).**

M. BUGAMELLI, F. SCHIVARDI and R. ZIZZA, *The euro and firm restructuring*, in A. Alesina e F. Giavazzi (eds): Europe and the Euro, Chicago, University of Chicago Press, **TD No. 716 (June 2009).**

B. HALL, F. LOTTI and J. MAIRESSE, *Innovation and productivity in SMEs: empirical evidence for Italy*, Small Business Economics, v. 33, 1, pp. 13-33, **TD No. 718 (June 2009).**

2010

A. PRATI and M. SBRACIA, *Uncertainty and currency crises: evidence from survey data*, Journal of Monetary Economics, v, 57, 6, pp. 668-681, **TD No. 446 (July 2002).**

L. MONTEFORTE and S. SIVIERO, *The Economic Consequences of Euro Area Modelling Shortcuts*, Applied Economics, v. 42, 19-21, pp. 2399-2415, **TD No. 458 (December 2002).**

S. MAGRI, *Debt maturity choice of nonpublic Italian firms* , Journal of Money, Credit, and Banking, v.42, 2-3, pp. 443-463, **TD No. 574 (January 2006).**

R. BRONZINI and P. PISELLI, *Determinants of long-run regional productivity with geographical spillovers: the role of R&D, human capital and public infrastructure*, Regional Science and Urban Economics, v. 39, 2, pp.187-199, **TD No. 597 (September 2006).**

E. IOSSA and G. PALUMBO, *Over-optimism and lender liability in the consumer credit market*, Oxford Economic Papers, v. 62, 2, pp. 374-394, **TD No. 598 (September 2006).**

S. NERI and A. NOBILI, *The transmission of US monetary policy to the euro area*, International Finance, v. 13, 1, pp. 55-78, **TD No. 606 (December 2006).**

F. ALTISSIMO, R. CRISTADORO, M. FORNI, M. LIPPI and G. VERONESE, *New Eurocoin: Tracking Economic Growth in Real Time*, Review of Economics and Statistics, v. 92, 4, pp. 1024-1034, **TD No. 631 (June 2007).**

A. CIARLONE, P. PISELLI and G. TREBESCHI, *Emerging Markets' Spreads and Global Financial Conditions*, Journal of International Financial Markets, Institutions & Money, v. 19, 2, pp. 222-239, **TD No. 637 (June 2007).**

U. ALBERTAZZI and L. GAMBACORTA, *Bank profitability and taxation*, Journal of Banking and Finance, v. 34, 11, pp. 2801-2810, **TD No. 649 (November 2007).**

M. IACOVIELLO and S. NERI, *Housing market spillovers: evidence from an estimated DSGE model*, American Economic Journal: Macroeconomics, v. 2, 2, pp. 125-164, **TD No. 659 (January 2008).**

F. BALASSONE, F. MAURA and S. ZOTTERI, *Cyclical asymmetry in fiscal variables in the EU*, Empirica, **TD No. 671**, v. 37, 4, pp. 381-402 **(June 2008).**

F. D'AMURI, O. GIANMARCO I.P. and P. GIOVANNI, *The labor market impact of immigration on the western german labor market in the 1990s*, European Economic Review, v. 54, 4, pp. 550-570, **TD No. 687 (August 2008).**

A. ACCETTURO, *Agglomeration and growth: the effects of commuting costs*, Papers in Regional Science, v. 89, 1, pp. 173-190, **TD No. 688 (September 2008).**

S. NOBILI and G. PALAZZO, *Explaining and forecasting bond risk premiums*, Financial Analysts Journal, v. 66, 4, pp. 67-82, **TD No. 689 (September 2008).**

A. B. ATKINSON and A. BRANDOLINI, *On analysing the world distribution of income*, World Bank Economic Review , v. 24, 1 , pp. 1-37, **TD No. 701 (January 2009).**

R. CAPPARIELLO and R. ZIZZA, *Dropping the Books and Working Off the Books*, Labour, v. 24, 2, pp. 139-162 ,**TD No. 702 (January 2009).**

C. NICOLETTI and C. RONDINELLI, *The (mis)specification of discrete duration models with unobserved heterogeneity: a Monte Carlo study*, Journal of Econometrics, v. 159, 1, pp. 1-13, **TD No. 705 (March 2009).**

L. FORNI, A. GERALI and M. PISANI, *Macroeconomic effects of greater competition in the service sector: the case of Italy*, Macroeconomic Dynamics, v. 14, 5, pp. 677-708, **TD No. 706 (March 2009).**

V. DI GIACINTO, G. MICUCCI and P. MONTANARO, *Dynamic macroeconomic effects of public capital: evidence from regional Italian data*, Giornale degli economisti e annali di economia, v. 69, 1, pp. 29-66, **TD No. 733 (November 2009).**

F. COLUMBA, L. GAMBACORTA and P. E. MISTRULLI, *Mutual Guarantee institutions and small business finance*, Journal of Financial Stability, v. 6, 1, pp. 45-54, **TD No. 735 (November 2009).**

A. GERALI, S. NERI, L. SESSA and F. M. SIGNORETTI, *Credit and banking in a DSGE model of the Euro Area*, Journal of Money, Credit and Banking, v. 42, 6, pp. 107-141, **TD No. 740 (January 2010).**

M. AFFINITO and E. TAGLIAFERRI, *Why do (or did?) banks securitize their loans? Evidence from Italy*, Journal of Financial Stability, v. 6, 4, pp. 189-202, **TD No. 741 (January 2010).**

S. FEDERICO, *Outsourcing versus integration at home or abroad and firm heterogeneity*, Empirica, v. 37, 1, pp. 47-63, **TD No. 742 (February 2010).**

V. DI GIACINTO, *On vector autoregressive modeling in space and time*, Journal of Geographical Systems, v. 12, 2, pp. 125-154, **TD No. 746 (February 2010).**

S. MOCETTI and C. PORELLO, *How does immigration affect native internal mobility? new evidence from Italy*, Regional Science and Urban Economics, v. 40, 6, pp. 427-439, **TD No. 748 (March 2010).**

A. DI CESARE and G. GUAZZAROTTI, *An analysis of the determinants of credit default swap spread changes before and during the subprime financial turmoil*, Journal of Current Issues in Finance, Business and Economics, v. 3, 4, pp., **TD No. 749 (March 2010).**

P. CIPOLLONE, P. MONTANARO and P. SESTITO, *Value-added measures in Italian high schools: problems and findings*, Giornale degli economisti e annali di economia, v. 69, 2, pp. 81-114, **TD No. 754 (March 2010).**

A. BRANDOLINI, S. MAGRI and T. M SMEEDING, *Asset-based measurement of poverty*, Journal of Policy Analysis and Management, v. 29, 2 , pp. 267-284, **TD No. 755 (March 2010).**

G. CAPPELLETTI, *A Note on rationalizability and restrictions on beliefs*, The B.E. Journal of Theoretical Economics, v. 10, 1, pp. 1-11,**TD No. 757 (April 2010).**

S. DI ADDARIO and D. VURI, *Entrepreneurship and market size. the case of young college graduates in Italy*, Labour Economics, v. 17, 5, pp. 848-858, **TD No. 775 (September 2010).**

A. CALZA and A. ZAGHINI, *Sectoral money demand and the great disinflation in the US*, Journal of Money, Credit, and Banking, v. 42, 8, pp. 1663-1678, **TD No. 785 (January 2011).**

2011

S. DI ADDARIO, *Job search in thick markets*, Journal of Urban Economics, v. 69, 3, pp. 303-318, **TD No. 605 (December 2006).**

E. CIAPANNA, *Directed matching with endogenous markov probability: clients or competitors?*, The RAND Journal of Economics, v. 42, 1, pp. 92-120, **TD No. 665 (April 2008).**

L. FORNI, A. GERALI and M. PISANI, *The Macroeconomics of Fiscal Consolidation in a Monetary Union: the Case of Italy*, in Luigi Paganetto (ed.), Recovery after the crisis. Perspectives and policies, VDM Verlag Dr. Muller, **TD No. 747 (March 2010).**

A. DI CESARE and G. GUAZZAROTTI, *An analysis of the determinants of credit default swap changes before and during the subprime financial turmoil*, in Barbara L. Campos and Janet P. Wilkins (eds.), The Financial Crisis: Issues in Business, Finance and Global Economics, New York, Nova Science Publishers, Inc., **TD No. 749 (March 2010).**

G. GRANDE and I. VISCO, *A public guarantee of a minimum return to defined contribution pension scheme members*, The Journal of Risk, v. 13, 3, pp. 3-43, **TD No. 762 (June 2010).**

P. DEL GIOVANE, G. ERAMO and A. NOBILI, *Disentangling demand and supply in credit developments: a survey-based analysis for Italy*, Journal of Banking and Finance, v. 35, 10, pp. 2719-2732, **TD No. 764 (June 2010).**

M. TABOGA, *Under/over-valuation of the stock market and cyclically adjusted earnings*, International Finance, v. 14, 1, pp. 135-164, **TD No. 780 (December 2010).**

S. NERI, *Housing, consumption and monetary policy: how different are the U.S. and the Euro area?*, Journal of Banking and Finance, v.35, 11, pp. 3019-3041, **TD No. 807 (April 2011).**

FORTHCOMING

M. BUGAMELLI and A. ROSOLIA, *Produttività e concorrenza estera*, Rivista di politica economica, **TD No. 578 (February 2006).**

G. DE BLASIO and G. NUZZO, *Historical traditions of civicness and local economic development*, Journal of Regional Science, **TD No. 591 (May 2006).**

F. CINGANO and A. ROSOLIA, *People I know: job search and social networks*, Journal of Labor Economics, **TD No. 600 (September 2006).**

F. SCHIVARDI and E. VIVIANO, *Entry barriers in retail trade*, Economic Journal, **TD No. 616 (February 2007).**

G. FERRERO, A. NOBILI and P. PASSIGLIA, *Assessing excess liquidity in the Euro Area: the role of sectoral distribution of money*, Applied Economics, **TD No. 627 (April 2007).**

P. E. MISTRULLI, *Assessing financial contagion in the interbank market: maximun entropy versus observed interbank lending patterns*, Journal of Banking & Finance, **TD No. 641 (September 2007).**

Y. ALTUNBAS, L. GAMBACORTA and D. MARQUÉS, *Securitisation and the bank lending channel*, European Economic Review, **TD No. 653 (November 2007).**

M. BUGAMELLI and F. PATERNÒ, *Output growth volatility and remittances*, Economica, **TD No. 673 (June 2008).**

V. DI GIACINTO e M. PAGNINI, *Local and global agglomeration patterns: two econometrics-based indicators*, Regional Science and Urban Economics, **TD No. 674 (June 2008).**

G. BARONE and F. CINGANO, *Service regulation and growth: evidence from OECD countries*, Economic Journal, **TD No. 675 (June 2008).**

S. MOCETTI, *Educational choices and the selection process before and after compulsory school*, Education Economics, **TD No. 691 (September 2008).**

P. SESTITO and E. VIVIANO, *Reservation wages: explaining some puzzling regional patterns*, Labour, **TD No. 696 (December 2008).**

P. PINOTTI, M. BIANCHI and P. BUONANNO, *Do immigrants cause crime?*, Journal of the European Economic Association, **TD No. 698 (December 2008).**

R. GIORDANO and P. TOMMASINO, *What determines debt intolerance? The role of political and monetary institutions*, European Journal of Political Economy, **TD No. 700 (January 2009).**

F. LIPPI and A. NOBILI, *Oil and the macroeconomy: a quantitative structural analysis*, Journal of European Economic Association, **TD No. 704 (March 2009).**

F. CINGANO and P. PINOTTI, *Politicians at work. The private returns and social costs of political connections*, Journal of the European Economic Association, **TD No. 709 (May 2009).**

Y. ALTUNBAS, L. GAMBACORTA, and D. MARQUÉS-IBÁÑEZ, *Bank risk and monetary policy*, Journal of Financial Stability, **TD No. 712 (May 2009).**

P. ANGELINI, A. NOBILI e C. PICILLO, *The interbank market after August 2007: What has changed, and why?*, Journal of Money, Credit and Banking, **TD No. 731 (October 2009).**

G. BARONE and S. MOCETTI, *Tax morale and public spending inefficiency*, International Tax and Public Finance, **TD No. 732 (November 2009).**

L. FORNI, A. GERALI and M. PISANI, *The macroeconomics of fiscal consolidations in euro area countries*, Journal of Economic Dynamics and Control, **TD No. 747 (March 2010).**

G. BARONE, R. FELICI and M. PAGNINI, *Switching costs in local credit markets*, International Journal of Industrial Organization, **TD No. 760 (June 2010).**

G. BARONE and S. MOCETTI, *With a little help from abroad: the effect of low-skilled immigration on the female labour supply*, Labour Economics, **TD No. 766 (July 2010).**

S. MAGRI and R. PICO, *The rise of risk-based pricing of mortgage interest rates in Italy*, Journal of Banking and Finance, **TD No. 778 (October 2010).**

A. ACCETTURO and G. DE BLASIO, *Policies for local development: an evaluation of Italy's "Patti Territoriali"*, Regional Science and Urban Economics, **TD No. 789 (January 2006).**

E. COCOZZA and P. PISELLI, *Testing for east-west contagion in the European banking sector during the financial crisis*, in R. Matoušek; D. Stavárek (eds.), Financial Integration in the European Union, Taylor & Francis, **TD No. 790 (February 2011).**

S. NERI and T. ROPELE, *Imperfect information, real-time data and monetary policy in the Euro area*, The Economic Journal, **TD No. 802 (March 2011).**